Network Models of the Lateral Intraparietal Area

Wujie Zhang

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Network Models of the Lateral Intraparietal Area

Wujie Zhang

The monkey lateral intraparietal area (LIP) is involved in visual attention and eye movements. It has traditionally been studied using extracellular recording, where often a single neuron is recorded at a time. Thus we have a wealth of correlational knowledge of what LIP neurons do, but not how or why, i.e. we do not know the circuit mechanisms and functions of the observed LIP activity. In this thesis, we have aimed to uncover the circuit mechanisms underlying LIP activity by building tightly constrained computational models.

In Part 1, we found that during two versions of a delayed-saccade task, beneath similar population average firing patterns across time lie radically different network dynamics. When neurons are not influenced by stimuli outside their receptive fields (RFs), dynamics of the high-dimensional LIP network lie predominantly in one multi-neuronal dimension, as predicted by an earlier model. However, when activity is suppressed by stimuli outside the RF, LIP dynamics markedly deviate from a single dimension. The conflicting results can be reconciled if two LIP local networks, each dominated by a single multi-neuronal activity pattern, are suppressively coupled to each other. These results demonstrate the low dimensionality of LIP local dynamics and suggest active involvement of LIP recurrent circuitry in surround suppression and, more generally, in processing attentional and movement priority and in related cognitive functions.

In Part 2, we examine the mechanisms of learning in LIP. When monkeys learn to group visual stimuli into arbitrary categories, LIP neurons become category-selective. Surprisingly, the representations of learned categories are overwhelmingly biased: while different categories are

behaviorally equivalent, nearly all LIP neurons in a given animal prefer the same category. We propose that Hebbian plasticity, at the synapses to LIP from prefrontal cortex and from lower sensory areas, could lead to the development of biased representations. In our model, LIP category selectivity arises due to competition between inputs encoding different categories, and bias develops due to excitatory lateral interactions among LIP neurons. This model reproduces the different levels of category selectivity and bias observed in multiple experiments. Our results suggest that the connectivity of LIP allows it to learn the behavioral importance of stimuli in order to guide attention.

Table of Contents

List of Figures

1 Coupling between one-dimensional networks reconciles conflicting dynamics in LIP and reveals its recurrent circuitry

1.1 Introduction

It has become increasingly appreciated that neural functions need to be understood in terms of neuronal populations and the dynamics of the circuits to which they belong (Miller and Wilson, 2008; Buzsaki, 2010; Shenoy et al., 2013). However, the field of systems neuroscience in nonhuman primates has been dominated by electrophysiology studies in which a single neuron or a few neurons are recorded at a time. Thus, while we have a wealth of knowledge of single neuron behaviors in many areas of the primate brain, this knowledge remains largely phenomenological—we know *what* neurons do, but not *how* they do it: especially on the circuit level, the mechanisms and connectivity underlying single neuron behaviors are often obscure.

Such is the case in LIP, where a large body of literature has revealed that the activity of single neurons encodes visual attention and saccadic eye movements, as well as decision making variables, abstract categories, and other cognitive variables (Andersen and Cui, 2009; Bisley and Goldberg, 2010; Freedman and Assad, 2011; Gold and Shadlen, 2007; Gottlieb, 2007; Kable and Glimcher, 2009). However, little is known about the circuitry inside or outside the LIP network that produces such activity, and therefore the role of LIP in many of these functions is controversial. A step in understanding this circuitry was taken by Ganguli et al. (2008), who analyzed LIP network dynamics during two different tasks: a delayed saccade task (Bisley and Goldberg, 2003, 2006) and a random-dot motion discrimination task (Roitman and

Shadlen, 2002). They found that the dynamics of the high-dimensional LIP network are dominated by one multi-neuronal dimension on slow timescales, and this one-dimensionality was key to explaining a nontrivial, unexpected correspondence between LIP single neuron responses and the timing of attentional shifts (examined in more detail below). More recently, Fitzgerald et al. (2013) found further evidence for one-dimensional dynamics in three experiments in which LIP encoded learned associations between visual stimuli.

Using a delayed saccade task similar to the task of Bisley and Goldberg (2003, 2006; hereafter BG), Falkner, Krishna et al. (2010; hereafter FK) reported "surround suppression" in LIP (see also Louie et al., 2011), a phenomenon seen in many visual areas during which stimuli outside the RF of a cell suppress the cell's activity (Allman et al., 1985; Sundberg et al., 2009; Tsui and Pack, 2011). In the FK study, the population mean activity pattern of LIP is very similar to that in the BG study, as expected given the very similar tasks. However, we find that the pattern of activity across neurons changes over time in a very different way in the FK study. In particular, the network dynamics in the FK dataset markedly deviate from the one-dimensional dynamics observed in the BG dataset, calling into question the validity of the one-dimensional LIP model of Ganguli et al (2008). We show that the two sets of conflicting results can be reconciled and well characterized by a more general low-dimensional model, in which each of two local LIP networks in isolation would have its own single dominant dimension, but suppressive coupling between them gives rise to two dominant multi-neuronal activity patterns. These patterns explain the observed dynamics and provide a mechanism for surround suppression. Finally, our modeling suggests that the network dynamics observed by FK

provide a signature indicating that surround suppression results from recurrent interactions within LIP, rather than being inherited from the inputs from other areas. This means that local recurrent processing in LIP contributes to computation of attentional priority and other decision variables. Our study thus represents a step forward in discovering the circuit mechanisms and connectivity from single neuron recordings, and paves the way for a mechanistic understanding of LIP functions.

1.2 Results

One-dimensional dynamics in LIP

We begin by describing the first of the two conflicting datasets (Bisley and Goldberg, 2003), along with the one-dimensional model (Ganguli et al., 2008) to which it gave rise.

The delayed saccade task of BG is described in Fig. 1.1A (details in Supplemental Information [SI] section 1). During this task, LIP neurons exhibit a large transient visual response to the onset of a saccade target or distractor in the RF, and sustained delay period activity (delay activity) when a saccade is planned to the RF (Fig. 1.1C). When a distractor is flashed away from the target location during the delay period, attention is transiently attracted away from the target location to the distractor location. At the same time, the average visual response level of LIP neurons whose RFs contain the distractor location (the distractor population) rises above the average delay activity level of neurons whose RFs contain the target location (the target population). As the visual activity of the distractor population decays back to baseline, the locus of attention shifts back to the target location. This shift in attention coincides with the shift in the peak of LIP activity

from the distractor population to the target population: when the decaying visual activity of the distractor population drops to a level statistically indistinguishable from the sustained delay activity of the target population (the "crossing time"—when the decaying red trace crosses the blue trace in Fig. 1.1C), neither the target nor the distractor location has attentional advantage, whereas 100-250 ms before or after this crossing time, the distractor or target location, respectively, is the clear locus of attention.

Further analyses of these results (Bisley and Goldberg, 2006) revealed that this correspondence between activity crossing and attentional switching also held at the level of single LIP neurons. The crossing time of a single neuron is defined as the time at which the neuron's decaying response to a distractor, on trials in which a distractor is in its RF (distractor trials), crosses its own level of delay activity on trials in which a target is in its RF (target trials). These single-neuron crossing times are surprisingly invariant across neurons and closely aligned with the monkey's attentional switching time, despite high variability across neurons in their peak visual responses, time constants of visual response decay, and delay period responses.

Ganguli et al. (2008) explained this observation with the proposal that the dynamics of a local network of LIP neurons are dominated on slow timescales by one multi-neuronal activity pattern (i.e., a pattern or vector of relative firing rates across the cells of the network). Briefly, the recurrent connectivity of a local network causes certain multi-neuronal activity patterns to persist longer in the absence of input; given steady input, these slowly decaying patterns also build up to be strongly amplified. If the network has only a single pattern that decays slowly, we refer to it as the network's "slow mode," where "mode" is a term borrowed from physics that describes a characteristic

pattern of a system's response. As the visual response to a distractor decays, it becomes dominated by this slow mode after all other patterns decay away. Because the slow mode is more strongly amplified than other patterns, it also dominates steady-state responses, such as delay activity and activity during the initial fixation before target onset (fixation activity). Thus, after the other patterns in the distractor response decay away, the decaying distractor activity and the ongoing delay activity are both dominated by the slow mode, meaning that the pattern of distractor activity across neurons is very nearly a scaled-up version of the delay activity pattern. At the crossing time, when the amplitudes of distractor and delay activity are the same, the distractor activity pattern is very nearly identical to the delay activity pattern. As a result, each individual neuron has roughly the same activity in its delay response as in its distractor response at the crossing time, so that all neurons have about the same single-neuron crossing time.

This one-dimensional model predicts that multi-neuronal activity patterns that change on slow timescales are all highly correlated with one another, because all are dominated by the same strongly amplified pattern. These include fixation and delay activity patterns, and, to a lesser extent, slowly decaying visual activity patterns. On the other hand, during the initial transient visual response, many other activity patterns are activated, so the transient visual activity pattern is not highly correlated with the steady-state activity patterns. Ganguli et al. (2008) confirmed these predictions using the following analysis, which reveals network dynamics from the activity of a population of singly recorded neurons. At any millisecond time point *t* during a trial, we represent the trial-averaged activity of a population of *n* neurons as an *n*-dimensional vector, $\vec{r}(t)$, in an *n*-dimensional multi-neuronal firing rate space; each of the *n* elements of $\vec{r}(t)$ is the

activity of one neuron at time *t*, averaged over trials. We also compute the *n*-dimensional fixation activity vector, $\vec{F}$, where each element is the activity of one neuron averaged over the fixation period before target onset and over target trials. Then, at each time point *t* over the course of the trial, a correlation coefficient is computed between $\vec{F}$ and $\vec{r}(t)$. Fig. 1.1E shows that the correlation to fixation activity is indeed high for delay activity or distractor activity after the transient visual response decays away, indicating that fixation, delay, and post-transient distractor activity patterns all lie roughly in a single dimension, corresponding to the dominant activity pattern. The drop in correlation coefficient during the visual response indicates the transient deviation of activity from this one dimension caused by the transient activation of other non-dominant patterns.

Surround suppression and violations of one-dimensional dynamics

We continue by describing the second of the two conflicting datasets (Falkner, Krishna et al., 2010) and how it exhibits large deviations on both fast and slow timescales from the predictions of the one-dimensional model.

The task of FK (Fig. 1.1B) is very similar to that of BG. For both tasks, we analyze data in each trial during time windows ending shortly after distractor onset (i.e., before the onset of the probe in the BG task; see Fig. 1.1A), up to which point the two tasks are virtually identical aside from three differences. First, BG used a flashed target while FK presented a target that stayed visible during the delay. This does not result in qualitatively different delay activity levels (compare delay activity between Fig. 1.1C and D), consistent with LIP encoding the attentional and saccadic priority of the target location regardless of the visibility of that target. Second, BG randomly interleaved target

trials and distractor trials, while FK presented target and distractor trials in blocks. Thus, in the FK experiment, on almost every trial the monkey had an expectation of where the target and distractor would be. This is reflected in higher anticipatory firing on target trials compared to distractor trials during the fixation period before target onset. The third difference is likely to be the key difference that led to different neural responses observed during the two tasks. In the BG task, the target and distractor are in opposite visual quadrants and equidistant from the fixation spot. In the FK task, in contrast, either the target or the distractor is in the RF of the cell being recorded in a given session, and the other stimulus is at the location eliciting maximum surround suppression of the recorded neuron. With this placement of stimuli, a saccade plan to the surround significantly suppressed the visual response to the distractor, while distractor appearance in the surround transiently and weakly, but significantly, suppressed delay activity during saccade planning (Fig. 1.1D; quantified in Falkner, Krishna et al., 2010). Surround suppression was not observed in the BG dataset (examined in Bisley and Goldberg, 2006), in which the stimulus locations were not selected for suppression. Other than the surround suppression of response amplitudes, the FK dataset displays the same overall pattern of fixation, visual, and delay activity as the BG dataset (compare Fig. 1.1C and D).

However, beneath this apparent similarity in population average activity, the network dynamics are radically different; moreover, the FK dynamics clearly violate the predictions of the one-dimensional model. Fig. 1.1F shows the result of the correlation analysis on neural activity from the FK experiment. Most strikingly, on distractor trials (red trace), even though the appearance of the target in the surround only minimally

affects the mean firing rate of the distractor population, target appearance causes a large, sustained drop in correlation, when the one-dimensional model would predict an unchanging and high level of correlation, as in Fig. 1.1E. Furthermore, the later appearance of the distractor in the RF causes a large, transient rise in correlation which subsequently returns to the steady low level present before distractor onset, when the one-dimensional model would predict the opposite change—a large and transient drop in correlation upon distractor onset, as in Fig. 1.1E . On target trials (blue traces), the difference is more subtle, with target onset evoking a small, sustained drop in correlation, similar to the sustained drop in the BG case, but without the initially larger transient decrease.

Note that in the BG dataset, the two trial types are randomly interleaved; thus, the monkey does not know the trial type during the initial fixation, and fixation activities are the same on the two trial types. In the FK dataset, however, fixation activities are different on the two trial types due to the block design. We chose to use the fixation activity on target trials as opposed to distractor trials to calculate correlations because it reveals salient patterns in the network dynamics. Using distractor trial fixation activity is another angle from which to examine the network dynamics that gives less informative results, i.e., correlations do not rise and drop saliently over time (Fig. 1.S1A).

Thus, the results of BG and of FK seem incompatible. The robust one-dimensional dynamics observed by Ganguli et al. (2008) in the data of Bisley and Goldberg (2003, 2006) require that the local anatomical connectivity of LIP selectively amplify only one multi-neuronal activity pattern. How can this same anatomical connectivity realize dynamics that deviate so far from the one dominant multi-neuronal

pattern that it so strongly amplifies?

Simple model of coupled local networks reconciles the results

We found the answer in a simple model of the interactions between two coupled LIP local networks. This model replicates the FK findings and yet reduces to the one-dimensional dynamics that characterize the BG findings when the two local networks are not coupled.

We model two local networks of LIP neurons, each composed of excitatory (E) and inhibitory (I) neurons that share an RF, with randomly distributed neuronal time constants (Fig. 1.2A and B; see Experimental Procedures for details of the model). Within each local network, connections are sparse and their strengths are randomly distributed. The mean synaptic strengths of excitatory and inhibitory connections are such that, when there is no connection between the two local networks, the recurrent connectivity within each isolated local network amplifies a single multi-neuronal activity pattern much more strongly than all other patterns, making it decay much more slowly than all other patterns. We design the network so that this slowly decaying pattern is a pattern of increased activity across almost all local network neurons; this is achieved by setting overall excitatory strength stronger than overall inhibitory strength within each local network. This dominance of excitation is consistent with evidence based on dendritic structure of increased connectivity between excitatory cells in LIP compared to primary sensory cortices (Elston and Rosa, 1997).

The LIP cortical surface contains rough topological maps of visual space (Blatt et al., 1990; Patel et al., 2010). Neurons sharing an RF, which are more likely to be located

close to each other on the cortical surface, make up a local network in our model. We model the connections of I cells to be restricted to the local network to which they belong, as inhibitory interneurons generally only make short-range projections, whereas E cells can potentially make long-range projections to the other local network. Since no significant interaction between RFs was observed in the BG dataset (quantified in Bisley and Goldberg, 2006), we infer that, for these RFs, the corresponding local networks are not directly connected (Fig. 1.2A). In contrast, by maximizing surround suppression, FK selected for RFs that did interact. Since the interaction observed was predominantly suppressive, it's likely that the excitatory connections from each local network are stronger to the I cells than to the E cells of the other local network. For simplicity, we model the across-network connections as being from the E cells of each local network to the other local network's I cells only, with sparse and random connectivity (Fig. 1.2B). Our results do not change if we include weaker across-network E-to-E connections (data not shown).

We use a standard linear firing rate model to simulate the experiments (Dayan and Abbott, 2005). The experiments involve a variety of sensory, motor, and cognitive processes that likely give rise to a variety of external inputs to LIP during a trial, which we model as the following four types. (1) *Fixation input*: spontaneous firing from the external input sources when there is no stimulus in or saccade plan to the RF, such as during the fixation period. (2) *Visual input*: bottom-up input to a local network when a visual stimulus is in the RF, which is strong upon stimulus onset and becomes weak as the stimulus is sustained. Visual input arrives from areas that could include V2, V3, V3A, V4, middle temporal area (MT), and inferotemporal cortex (Baizer et al., 1991; Blatt et

al., 1990; Lewis and Van Essen, 2000). (3) *Delay input*: persistent top-down input to a local network when a saccade is being planned to the RF, arriving from frontal areas such as the frontal eye field (FEF) or dorsolateral prefrontal cortex (dlPFC; Blatt et al., 1990; Stanton et al., 1995; in SI section 8 we discuss other possible mechanisms underlying delay activity and their implications for our model). (4) *Expectation input*: top-down input to one local network during the fixation period before target onset, when the animal is in a block of trials during which the target always appears in the RF of that local network (as in the blocked experiment of FK). Expectation input likely also arrives from frontal areas such as FEF or dlPFC (Coe et al., 2002; Roesch and Olson, 2003). The total external input to the neurons at any time is the sum of one or more of these four kinds of input. For each of the four kinds of input, input to each cell is independently drawn from a uniform distribution, with ranges of the distributions chosen to fit experimentally observed neural responses. Thus, importantly, the inputs from different sources are uncorrelated. In addition, the external input contains weak, temporally correlated noise that is independent for different neurons, simply to produce small firing rate fluctuations similar to those seen in the experiments.

To simulate the single cell recording experiments, we run the simulation multiple times, each time with a different random instantiation of network connectivity, neuronal time constants, and input patterns, and "record" from a single randomly chosen cell during each simulation. Each simulation includes target and distractor trials for the recorded cell. Thus our simulated LIP population, like the experimental population, consists of single cells recorded at different times from different local LIP circuits. Fig. 1.2C and D show the population PSTHs from simulations of the BG (Fig. 1.2C) and the

FK (Fig. 1.2D) experiments, which reproduce the experimentally observed firing patterns, including the observed absence or presence of surround interactions. More significantly, our model reproduces the apparently conflicting network dynamics of the two experiments as revealed from the correlation analysis: the model of the BG experiment shows one-dimensional dynamics on slow timescales (Fig. 1.2E), and the model of the FK experiment shows the same higher-dimensional dynamics as experimentally observed (Fig. 1.2F).

If we compute correlations of instantaneous activity to distractor trial fixation activity, rather than to target trial fixation activity, the model also qualitatively reproduces the experimental results (Fig. 1.S1B). Furthermore, modeling higher reward levels as resulting in higher levels of delay input (Leon and Shadlen, 1999; Kennerly and Wallis, 2009), we reproduce the results found when the data of FK, which consists of trials with large or small reward, are analyzed separately by reward level (Fig. 1.S2). Because the activity and correlation patterns are qualitatively similar across reward levels in the data (Fig. 1.S2A-D), in all other simulations we simply modeled the average reward level.

Conceptual picture: coupling of local slow modes explain LIP dynamics

We fully analyze the behaviors of the model in the next sections, but first, in this section, we presage those results by presenting the simple conceptual understanding we arrived at through study of the model.

The answer to reconciling the two sets of results, slightly simplified, is the following. Each local network has its own single dominant activity pattern (its slow

mode), and therefore each on its own would follow one-dimensional dynamics. However, the circuitry that creates surround suppression causes these two patterns to suppress one another, and this mutual suppression in turn qualitatively explains the FK correlation patterns, as follows.

Suppose we are recording in one of the local networks, call it network 1, and let the other local network be network 2. $\vec{F}$, the fixation activity of network 1 on target trials, is dominated by its slow mode, being driven by both fixation input and expectation input. At other times, the correlation of network 1's instantaneous activity with $\vec{F}$ is high or low according to whether or not that instantaneous activity is dominated by the slow mode. Now consider network 1 on distractor trials. During the initial fixation period, network 1 receives fixation input but not expectation input. Thus, its slow mode is activated less than on target trials; in addition, its slow mode is suppressed by the more activated slow mode of network 2, which is receiving both fixation and expectation input. As a result, the relative contribution of activity patterns other than the slow mode to network 1's activity is larger than on target trials, resulting in reduced correlation between distractor trial fixation activity and $\vec{F}$. After the target appears in network 2's RF, network 1's slow mode continues to be driven only by fixation input; in addition, it is strongly suppressed by the slow mode of network 2, which is strongly driven by both visual stimulation and the subsequent top-down delay input. This greatly reduces the correlation. Finally, when the distractor appears, strong visual stimulation transiently drives up network 1's slow mode, which causes the transient rise in correlation.

The conceptual picture just given is simplified in that it describes each local network as having only one dimension of activity that is strongly amplified. In reality,

while each local network has only one strongly amplified dimension when it is isolated, a second strongly amplified dimension is created in each local network by the coupling between the two local networks. When network 2 is more strongly driven than network 1, not only is network 1's slow mode suppressed, but activity is also driven in network 1's second strongly amplified dimension, making the slow mode an even less dominant part of network 1's activity. This will become clear with the detailed analysis below.

Detailed analysis: two-dimensional dynamics result from the coupling of local slow modes

We now take a closer look at the mechanisms of the model. We modeled the BG scenario with two unconnected local networks, each having a dominant activity pattern, the slow mode. The model simply behaves like two copies of the one-dimensional model of Ganguli et al., reproducing one-dimensional dynamics and the absence of surround interaction.

The only difference in network architecture in our model of the FK scenario is the presence of connections between the two local networks. Thus, the dominant activity patterns of the two local networks influence each other and are no longer independent. To understand the activity patterns of the global network consisting of the two coupled local networks, we examine the global connectivity matrix, which describes all of the global network's connections, both within and between the local networks. This connectivity between neurons determines how strongly one neuron excites or inhibits other neurons—it can be equivalently described as connections between sets of activity patterns, determining how strongly activity in one pattern excites or inhibits activity in itself and in

other patterns. For a network composed of separate excitatory and inhibitory neurons, it is often informative to analyze its connectivity as the connections between its Schur activity patterns (Murphy and Miller, 2009; Goldman, 2009; described in more detail in SI section 2). These are an ordered set of orthogonal activity patterns whose connectivity with each other are as simple as possible for a set of orthogonal patterns: each Schur pattern has a self-connection, and in addition there is a set of purely feedforward connections between the patterns. We choose to order the patterns by their self-connection strength. Then, activity in pattern 1 (the pattern with the strongest self-excitation) can only influence itself by its self-connection; activity in pattern 2 can excite or inhibit activity in pattern 1, in addition to influencing itself; pattern 3 can excite or inhibit pattern 1, pattern 2, and itself, etc. Thus, given similar external inputs to the patterns, the dominance of any pattern in the network's dynamics can be predicted by the strengths of its self-connection and the feedforward connections it receives. The activity of the network at any moment can be uniquely decomposed as a weighted sum of all Schur patterns, and patterns that dominate would have weights with large absolute values (the weights can be positive or negative).

A set of numbers called eigenvalues can be calculated from the connectivity matrix; each eigenvalue is associated with a Schur pattern, and the real part of the eigenvalue corresponds to the strength of that pattern's self-connection. Plotting the eigenvalues of the global connectivity matrix from one representative simulation, we see that two eigenvalues have real parts more positive than the rest, indicating that there are two strongly self-excitatory activity patterns (Fig. 1.3A). Analysis of the feedforward connections between patterns shows that these two patterns are nearly independent, with

only a very weak feedforward connection from one to the other; furthermore, feedforward connections originating from the less self-excitatory patterns activate the two strongly self-excitatory patterns much more than they activate the less self-excitatory patterns (SI section 2 and Fig. 1.S3E and G).

To understand the structure of these two potentially dominant activity patterns, in Fig. 1.3B we plot the relative activation of different neurons in these two Schur patterns. We have arbitrarily chosen the overall sign of each pattern in Fig. 1.3B such that both have mostly positive elements in network 1, and we have arbitrarily set the amplitude (i.e. the vector norm) of each pattern to 1. We note two key points about these two global patterns, which together show that they represent the coupled activation of the two local slow modes. (1) The two patterns represent two different forms of coupled activation of the two local networks: one is a "sum pattern," representing roughly equal activation of the two local networks; the other is a "difference pattern," representing differential activation of the two local networks, i.e., this pattern increases the activity of one local network and decreases the activity of the other. (2) The portions of the sum and difference patterns within a given local network are very similar to each other (e.g., in Fig. 1.3B, compare the two patterns restricted to neurons 1-100; for this comparison, the overall sign of activation within a local network is arbitrary), as well as to the slow mode of that local network if it were not connected with the other local network (quantified in Fig. 1.S3H and SI section 4), which reflects the connectivity within that local network.

The sum and difference patterns, in addition to being strongly amplified by recurrent connectivity, also typically receive stronger external input than the other patterns. Consider the vector of external inputs $\vec{I}$, each of whose elements is the input to

one neuron of the global network. Let's decompose it as $\vec{I} = \vec{I}_{mean} + \vec{I}_{res}$, where the network 1 elements of $\vec{I}_{mean}$ all equal the mean input to network 1, which we call I_1, and similarly the network 2 elements are all I_2. $\vec{I}_{res}$ contains the residuals which sum to zero. Similarly we can decompose a Schur pattern $\vec{P}$ as $\vec{P} = \vec{P}_{mean} + \vec{P}_{res}$, where the elements of $\vec{P}_{mean}$ are P_1 and P_2, the local network means of $\vec{P}$. The external input to Schur pattern $\vec{P}$ is given by $\vec{I} \cdot \vec{P}$. Because over each local network, the residuals sum to zero while the mean vectors are constant, the dot product of any residual vector with any mean vector is 0. Thus $\vec{I} \cdot \vec{P} = \vec{I}_{mean} \cdot \vec{P}_{mean} + \vec{I}_{res} \cdot \vec{P}_{res}$. The first term $\vec{I}_{mean} \cdot \vec{P}_{mean} = N(I_1 P_1 + I_2 P_2)$, where N is the number of neurons in a local network. The second term is a dot product of uncorrelated random vectors. By the central limit theorem, for large N, $\vec{I}_{res} \cdot \vec{P}_{res}$ across different random instantiations of networks and inputs approaches a Gaussian distribution with mean zero and standard deviation $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$, where σ_I and σ_P are the standard deviations across the elements of $\vec{I}_{res}$ and $\vec{P}_{res}$, respectively. Thus, the typical order of magnitude of $\vec{I}_{res} \cdot \vec{P}_{res}$ will be $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$. To compare the magnitude of $N(I_1 P_1 + I_2 P_2)$ and $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$, we note that N is much greater than $\sqrt{N}$, and I_1 and I_2 are larger or comparable to σ_I (since external inputs are carried by purely excitatory projection neurons and thus are positive). On the other hand, for the sum and difference patterns, P_1 and P_2 are comparable to σ_P (Fig. 1.S6). Thus, the inputs to these two patterns are large and approximately $N(I_1 P_1 + I_2 P_2)$—intuitively, these two patterns represent concerted

activation of cells in each local network, and are thus driven by the mean inputs to each local network consistently across simulations. For the other patterns, P_1 and P_2 are close to zero and much smaller than σ_P (Fig. 1.S6). Thus, their inputs are small and dominated by $\sqrt{N}\sqrt{2}\sigma_I\sigma_P$—intuitively, they represent random activations of cells, and are weakly driven by the random fluctuations of inputs about their mean.

We note that for a small proportion of random instantiations of connectivity matrices, a pair of complex patterns (which are complex conjugates in the eigenvector basis) take the place of the single real global sum pattern described above. We show in SI section 5 and Fig. 1.S3H-L that, in these cases, the complex pattern pair behaves effectively like the single global sum pattern.

We can use our understanding of the two dominant global activity patterns to understand the activity within a single local network. Let's choose one of the local networks to be network 1. We will call the network 1 portions of the sum and difference patterns $\vec{S}\,1$ and $\vec{D}\,1$, respectively, and take them to be normalized to unit vector length. Because these two patterns are not exactly equal to one another, they define a two-dimensional space of strongly amplified activity patterns in network 1. A convenient orthogonal pair of vectors to serve as a basis for this space is a vector $\vec{a}\,1$ proportional to the average of $\vec{S}\,1$ and $\vec{D}\,1$, and a vector $\vec{d}\,1$ proportional to their difference (again, both normalized to unit vector length; Fig. 1.3C). $\vec{a}\,1$ is almost precisely the slow mode of the isolated network 1, while $\vec{d}\,1$ is very nearly orthogonal to that slow mode (see Fig. 1.S3H and SI section 4). From the above analysis, the activation of $\vec{S}\,1$ and $\vec{D}\,1$ is largely

determined by the mean inputs to the two local networks (Fig. 1.3D-E).

Detailed analysis: two-dimensional dynamics explain correlation patterns

We are now in a position to understand the behavior of correlations between fixation and instantaneous activities in the FK model. We begin by considering a population of neurons simultaneously recorded from a single local network, part of a single global network (Fig. 1.4). We then explain why the conclusions we reach remain valid for a population in which each neuron is recorded from a different global network (the case of our main simulations in Fig. 1.2 and likely of the FK experiment).

First, we see in simulations of a single global network that, indeed, the two dominant activity patterns, $\vec{S}\,1$ and $\vec{D}\,1$, largely explain the population-averaged activity of network 1 (Fig. 1.4A and E; the results and analysis are identical for network 2). Moreover, we can see the contributions of $\vec{S}\,1$ and $\vec{D}\,1$ activity to the correlation patterns by breaking up the correlations into two components, the component due to activity in the $\vec{S}\,1$ and $\vec{D}\,1$ patterns alone and the residual component, as follows. At any given time point, the correlation between instantaneous activity and fixation activity is $\frac{\hat{r}\cdot\hat{F}}{|\hat{r}||\hat{F}|}$, where $\hat{r}$ is the vector of mean-subtracted instantaneous activities (each element of $\hat{r}$ is the instantaneous activity of one neuron minus the population mean instantaneous activity), $\hat{F}$ is the vector of mean-subtracted fixation activities (each element of $\hat{F}$ is the fixation activity of one neuron minus the population mean fixation activity), and $|\cdot|$

denotes vector norm. We break $\hat{r}$ into components $\hat{r}_{sum}$, $\hat{r}_{diff}$, and $\hat{r}_{weak}$, the mean-subtracted instantaneous activity in the $\vec{S}1$ pattern, the $\vec{D}1$ pattern, and all other patterns, respectively, and do likewise for $\hat{F}$:

$$\begin{aligned}
\frac{\hat{r}\cdot\hat{F}}{|\hat{r}||\hat{F}|} &= \frac{(\hat{r}_{sum}+\hat{r}_{diff}+\hat{r}_{weak})\cdot(\hat{F}_{sum}+\hat{F}_{diff}+\hat{F}_{weak})}{|\hat{r}||\hat{F}|} \\
&= \frac{(\hat{r}_{sum}+\hat{r}_{diff})\cdot(\hat{F}_{sum}+\hat{F}_{diff})}{|\hat{r}||\hat{F}|} + \frac{(\hat{r}_{sum}+\hat{r}_{diff})\cdot\hat{F}_{weak}+\hat{r}_{weak}\cdot(\hat{F}_{sum}+\hat{F}_{diff})+\hat{r}_{weak}\cdot\hat{F}_{weak}}{|\hat{r}||\hat{F}|} \\
&= Corr_{sum,diff}+Corr_{residual}
\end{aligned}$$

The two terms $Corr_{sum,diff}$ and $Corr_{residual}$ that sum to the actual correlation are plotted in Fig. 1.4B and F—we see that $\vec{S}1$ and $\vec{D}1$ activity largely explains the qualitative changes in correlations over time.

Thus, the actual activity pattern across cells of the local network, $\vec{r}$, can be approximated as the vector sum of $\vec{S}1$ and $\vec{D}1$ activity, which determines the correlation patterns. Fig. 1.4C and G and Fig. 1.S4 illustrate $\vec{S}1$ and $\vec{D}1$ activity evolving over four time periods during a trial, as well as how their dynamics explain the correlation patterns.

Now we turn to examine how the dynamics of $\vec{S}1$ and $\vec{D}1$ activity are determined by their inputs. We have shown above that the input to the sum or different pattern is approximately $N(I_1 P_1+I_2 P_2)$. We note that the absolute values of P_1 and P_2 for the sum and difference patterns are all about equal (Fig. 1.3B), which we can call m. That is, for the sum pattern, $P_1 \approx P_2 \approx m$, while for the difference pattern, $P_1 \approx m$ and $P_2 \approx -m$. Thus, the inputs to $\vec{S}1$ and $\vec{D}1$ are approximately $Nm(I_1+I_2)$ and $Nm(I_1-I_2)$, respectively. When

the network is in a steady state, these inputs are amplified by the connectivity: $\vec{S}\,1$ and $\vec{D}\,1$ activity are given by $Nm(I_1+I_2)/(1-\lambda_S)$ and $Nm(I_1-I_2)/(1-\lambda_D)$ respectively, where λ_S and λ_D are the eigenvalues of the sum and difference patterns, respectively. As N, m, λ_S, and λ_D are all fixed properties of the network, the dynamics of $\vec{S}\,1$ and $\vec{D}\,1$ activity just depend on the dynamics of the mean inputs I_1 and I_2, being simply proportional to I_1+I_2 and I_1-I_2, respectively (Fig. 1.4C, D, G, H).

There are two finer points regarding this analysis and Fig. 1.4. First, note that, for the same time period, I_1 and I_2 are simply exchanged in distractor trials compared to target trials. Our approximation would thus predict that the two trial types would have the same magnitude and sign of $\vec{S}\,1$ activity, and the same magnitude and opposite sign of $\vec{D}\,1$ activity. However, the residual, stochastic part of the inputs to the two local networks are not simply exchanged on the two trial types, and their stochastic activations of $\vec{S}\,1$ and $\vec{D}\,1$ result in different vector lengths for the same time period in Fig. 1.4C compared to Fig. 1.4G. For the same reason, during the transient response at time (3) in Fig. 1.4C and G, the particular random instantiation of stochastic inputs in that simulation happens to make the small $\vec{D}\,1$ activity point in the same direction for both trial types,.

Second, one might expect transient visual stimulation to activate non-dominant patterns and bring correlations to the same low level in both the BG and FK cases. However, during both the target and distractor visual responses, the correlation is higher in the FK data than in the BG data (compare Fig. 1.1E and F), which we reproduce in our model (compare Fig. 1.2E and F). We discuss in SI section 6 how this correlation

difference might result from a difference in the variability of visual inputs between BG and FK.

Unconnected neurons behave like neurons in a single local network

In Fig. 1.4, we modeled the results of recording from a neuronal population belonging to a single local network. In our main simulations (Fig. 1.2), we instead reproduce the experimental procedure, by modeling cells recorded during different experimental sessions as coming from independent sub-networks of LIP, i.e., from independent random instantiations of the global network and its inputs. However, the above analysis still applies.

A neuron tends to have similar activation in its network's sum and difference patterns; this activation is determined by the particular instantiation of the probabilistic connectivity. Now consider the "virtual" dominant sum or difference pattern of a population of neurons from different networks, determined by setting each neuron's activity to its activity in its own network's sum or difference pattern, respectively. Although external inputs to individual cells are variable and noisy across networks and sessions, the sum or difference patterns of each network, and thus the dominant virtual patterns, are primarily driven by the mean inputs across local networks, which are consistent across networks and sessions. Therefore the virtual dominant patterns are activated in roughly the same manner during a trial as the dominant patterns of a single network.

Then, during steady-state activity (i.e., fixation activity and delay activity) the correlation pattern of the population drawn from different networks behaves in the same

way as a population from a single network. Outside of the steady states (i.e., transient visual activity), the activations of the virtual dominant patterns are consistent with activations of the actual dominant patterns, as long as the actual dominant patterns of different networks have similar time constants. These time constants are determined by the neuronal time constants as well as the eigenvalues and other properties of the connectivity within a given network, with the dominant eigenvalues largely determined by the mean connection strengths within and between E and I populations. Because we model different local networks as having the same statistics of neuronal time constants and connectivity parameters, we expect the time constants of the dominant patterns to be reasonably similar across local networks (see the Supplemental Data of Ganguli et al., 2008, which shows the invariance across local networks of the local slow mode decay time). We found that in our model, within a robust range of the variability of these parameters, correlation during transient states as well as steady states is indeed similar between a population of neurons drawn from different networks and a population drawn from a single network.

Direct evidence for two-dimensional dynamics in the Falkner, Krishna et al. dataset

Since we propose that the BG data is dominantly one-dimensional and the FK data two-dimensional, we used principal component analysis (PCA) to directly examine the dimensionality of the two datasets. We focus on distractor trials because our correlation analysis revealed that they show the most salient dynamical differences between BG and FK. We excluded the transient visual responses to the distractor as they involve activation of weak patterns, and did PCA on the remaining slow timescales

activity patterns. The results indeed confirm the one-dimensionality of BG and two-dimensionality of FK (Fig. 1.5A-B).

Given the 2D space spanned by the top two principal components (PCs) identified from the FK data, we ask further, do activity patterns in this dominant 2D space actually behave as our model predicts? To answer this question, we first estimate the activity directions in the data that correspond to the ones in our model. We cannot estimate the directions of $\vec{S}1$ and $\vec{D}1$, but activity in those directions can be equivalently described as activity in the directions of $\vec{a}1$ and $\vec{d}1$. We can assume that the direction having the maximum mean firing rate within the 2D space of the 2 PCs is close to the direction of $\vec{a}1$, since $\vec{a}1$ is a direction representing concerted firing of neurons in a local network, thus arriving at the putative $\vec{a}1$ and $\vec{d}1$ of the data (Fig. 1.5C). In Fig. 1.5D-G we plot the activations over time of $\vec{a}1$ and $\vec{d}1$ from data and model. The activations of the putative $\vec{a}1$ and $\vec{d}1$ of the data match those predicted by the model, providing direct evidence that our proposed two-dimensional dynamics underlie the FK data.

Two-dimensional dynamics reveal the recurrent origin of LIP surround suppression

The phenomenon of surround suppression is observed in multiple cortical areas (reviewed in Rubin et al., 2015) and has been extensively studied as a model for understanding cortical computations and circuit mechanisms (e.g. in V1, Ozeki et al., 2009, Rubin et al., 2015, and see review in Nurminen and Angelucci, 2014). When considering surround suppression in a given cortical area, a key mechanistic question is

the following: to what extent is the suppression inherited from surround suppression in other areas, i.e. resulting from a withdrawal of input from those areas, and to what extent is it due to reciprocal, suppressive coupling within the area between regions that respond to the center and surround stimuli? Such coupling might be mediated by within-area horizontal connections and/or by projections to and from other areas (involving "feedback" projections, e.g., see Angelucci and Bressloff, 2006), but in either case we shall refer to such reciprocal coupling as "recurrent."

Of areas that directly or indirectly project to LIP (Blatt et al., 1990; Clower et al., 2001), surround suppression has been observed in MT (Born and Bradley, 2005; Hunter and Born, 2011; Tsui and Pack, 2011), V4 (Desimone and Schein, 1987; Schein and Desimone, 1990; Sundberg et al., 2009), superior colliculus (Dorris et al., 2007), FEF (Schall and Hanes, 1993; Schall et al., 1995; Cavanaugh et al., 2012), and dlPFC (Suzuki and Gottlieb, 2013). Thus, it is possible that LIP surround suppression is inherited from the inputs to LIP from one or more of these areas. However, according to our model, the observed pattern of correlation between fixation and instantaneous activity depends crucially on activity patterns arising from the coupling of local LIP networks. We argue that the experimentally observed correlation pattern is a signature indicating that withdrawal of external input, or more generally any alteration of external input to a single uncoupled local network, cannot account for LIP surround suppression. Instead, mutual interactions between local networks within LIP play a key role.

This can be demonstrated by simulating the scenario of the null hypothesis—LIP surround suppression being inherited from external inputs. In this version of the model, the two local networks are uncoupled. Whenever a stimulus appears or a saccade is

planned, the local network with the corresponding RF is activated by visual or delay input; at the same time, the external input to the other local network is reduced, modeling surround suppression inherent in one or more input sources (see Experimental Procedures for model details). Fig. 1.6A shows the population average PSTHs from a simulation of the FK experiment using this model. On the surface, if we examine only the firing rates, this model of surround suppression reproduces the experimental data. However, if we examine the underlying network dynamics using the correlation analysis (Fig. 1.6B), we find that this model cannot reproduce the experimentally observed correlation pattern. Specifically, the dynamics of each local network here is dominated by its slow mode, more similar to the BG dataset (Fig. 1.1E and 1.2E). In Fig. 1.S5 and SI section 7, we discuss and rule out other forms of the null hypothesis, including the possibility that the network dynamics we postulate are in another area from which LIP inherits its correlations. We conclude, therefore, that the most likely and parsimonious interpretation is that surround suppression in LIP arises from its internal circuitry.

The consequences of low-dimensional dynamics for attentional switching

As described above in the section "One-dimensional dynamics in LIP," BG found that the crossing times of LIP single neurons coincided with the monkey's attentional switching time, the time it takes to switch attention from the distractor back to the target. The crossing times of single neurons are post hoc observations not available on single trials and do not have a causal neural function, but they are signatures of slow mode dynamics that are at work on single trials and allow invariant behavioral switching times across trials and spatial locations despite heterogeneous single neuron properties.

As we've shown, in the FK condition the dynamics of an LIP local network are no longer restricted to one dimension. LIP single neurons having a common crossing time depends on one-dimensional dynamics: the slowly-decaying population visual response to the distractor and the population delay activity lying on a single dimension (Ganguli et al., 2008). In FK, both the visual response and the delay activity evolve in a two-dimensional space, and there is no longer guarantee that the two should meet in state space. Thus, our model predicts that LIP single neurons would no longer have a common crossing time when visual stimuli interact as in FK (Fig. 1.S7C and SI section 9). Indeed, that is what we found in the FK data (Fig. 1.S7E). The stochastic connectivity and inputs that contribute to this variance of the single neuron crossing times are likely to also result in variance of population crossing times across trials and spatial locations (SI section 9). Thus, to the extent that LIP causally mediate attentional allocation, we predict that a monkey's attentional switching time would be more variable across trials and spatial locations under the FK condition than the BG condition. The variability across spatial locations can be tested using the psychophysical methods of BG.

Network dynamics underlying different levels of surround suppression

While mapping the visual location of maximum surround suppression for a given RF, FK ran the task with one stimulus (target or distractor) at the RF, and the other stimulus at a variety of locations in the surround that elicited varying levels of suppression. Because of the small number of trials at locations other than the maximum suppression location, we could not reliably calculate correlations at these locations. Thus we studied the network dynamics underlying different levels of suppression using our

model, by modeling pairs of local networks with different across-local-network E-to-I synaptic weights. First, we see that as these weights increase, from the BG case of no connection to the case of maximum suppression in FK, the two independent slow modes of the two local networks gradually morph into the sum and difference patterns coupling the two local networks (Fig. 1.S8). As the dominant activity patterns of the network gradually change, we expect them to lead to gradual changes in the correlation patterns. Fig. 1.7 shows our model predictions for correlation patterns at intermediate levels of suppression, where we've focused on the correlations on distractor trials because they show the most salient changes from the BG to the FK case. In particular, as coupling between the local networks increases, the steady-state correlation during the delay period deceases, and the drop in correlation upon distractor onset becomes smaller and eventually turn into a rise in correlation. These effects are due to the gradual emergence of the dominant difference pattern. As the number of neurons that can be simultaneously recorded from LIP increases in the future, these predictions will become easier to test, since each visual location would elicit different levels of surround suppression for different neurons.

1.3 Discussion

By uncovering a recurrent mechanism likely to underlie LIP surround suppression, our study suggests the active involvement of LIP in attentional and saccadic selection, and in perceptual and value-based oculomotor decision making, cognitive processes in which the active role of LIP has often been debated. LIP is part of a fronto-parietal-collicular network that mediates attentional guidance and eye movements, and

the attentional and saccadic priorities associated with different locations (a "priority map") are encoded in the activity of neurons in this network with the corresponding RF locations (Bisley and Goldberg, 2010; Andersen and Cui, 2009). It has long been theorized that different locations on this priority map mutually suppress each other to facilitate attentional and saccadic selection, to allow persistent focus by resisting distraction, and to allow the planning and execution of sequential saccades (Itti and Koch, 2001; Constantinidis and Wang, 2004; Xing and Andersen, 2000). However, the neural substrates and mechanisms of these processes is not clear. Our results suggest that LIP directly participates in these processes and shapes the priority map, instead of merely reflecting computations achieved in other areas. One specific attentional phenomenon for which our results provide a potential mechanism is the "set-size effect": during visual search, it takes longer to find a target when there are more distractors, and, correspondingly, LIP activity is lower when there are more distractors (Balan et al., 2008). This may be the result of increased surround suppression by larger numbers of activated LIP local networks, which should yield correlations corresponding to a higher-dimensional dominant activity space (e.g., number of dimensions equal to the number of mutually interacting networks). Our study thus provides a basis for analyzing activity dynamics when multiple stimuli evoke interaction of multiple local LIP networks, as occurs in natural visual environments.

Our results have implications for the mechanisms underlying certain types of perceptual decision making, where saccadic decisions are made based on noisy sensory evidence (Roitman and Shadlen, 2002; Gold and Shadlen, 2007). This type of decision making has been posited to involve two neuronal pools that integrate opposing sensory

evidence, which either each accumulates evidence independently and races toward a decision (Mazurek et al., 2003), or compete with each other by mutual inhibition (Wong and Wang, 2006; Usher and McClelland, 2001). To the extent that such neuronal pools are in LIP, which of the two classes of models applies to each instance of decision making would depend on whether the two neuronal pools are recurrently coupled. When they are not coupled (like the neuronal pools studied by BG), the independent accumulator model would apply, and when they are coupled, the mechanisms described here would contribute to the competition that leads to decision making. In the future it would be interesting to study theses two cases of decision making separately (as BG and FK have done in studying attentional switching), examine the neural correlates on the population level, and compare them with behavior.

These LIP interactions would cause priority assignments to be in part determined in relative terms, as has been observed in certain forms of value-based decision making. In saccade tasks where different saccade targets are associated with different magnitudes of reward or reward probabilities, some LIP neurons encode the expected value of different saccades (Platt and Glimcher, 1999; Dorris and Glimcher, 2004). Importantly, the value representation in LIP is relative, such that the response to one saccade target depends on its value relative to those of other possible saccade targets; this relative value encoding is well described by the phenomenological model of divisive normalization (Louie et al., 2011; Carandini and Heeger, 2012). Surround suppression, computed within LIP in ways similar to those described here, provides a circuit mechanism for divisive normalization of value representations (Louie et al., 2014; LoFaro et al., 2014; Rubin et al., 2015).

Regardless of the cognitive context in which LIP function has been investigated, research has often focused on single neuron activity or the average activity of LIP populations. Our work adds to other recent work (e.g., Churchland et al., 2012; Cunningham and Yu, 2014) in suggesting that there is much information in the activity patterns across neurons, which change as a function of external stimuli and internal goals such as saccade plans. As we have seen, even when the mean activity of a population changes only subtly, the pattern of activity across neurons can change drastically (e.g., when a target appears outside the RF of a local LIP population). Thus, beyond the information carried by single neurons or their average activity, downstream areas could potentially read out information from the activity pattern across LIP neurons—although, whether or how downstream areas do this remains to be tested. This is potentially important in the natural context, where LIP local networks must interact to process a multitude of changing visual stimuli and internal goals to guide visuomotor behavior.

1.4 Methods

Data processing

To estimate the standard error of correlations between instantaneous and fixation activities from an actual population or a simulated population, we formed 1000 bootstrap sample populations by sampling cells with replacement from the given population, and computed standard errors from the correlations calculated from each bootstrap sample population.

Modeling details

The model network consists of two local networks of N neurons each (N/2 E cells and N/2 I cells). We included I cells unlike the E-cells-only model of Ganguli et al. (2008) because we aimed to model surround suppression. We chose to model equal numbers of E and I cells for simplicity, but modeling more realistic ratios of the number of E and I cells does not change our results (data not shown). Within a local network, the mean excitatory and inhibitory synaptic weights, onto both E and I cells, are $\frac{a}{N/2}$ and $-\frac{b}{N/2}$, respectively. We choose $a > 1$ and $a - b < 1$, such that each local network operates as an inhibition-stabilized network, a network regime underlying surround suppression in V1 (Ozeki et al., 2009; Rubin et al., 2015). Furthermore, $a > b$, so that each local network strongly amplifies a pattern of increased activity across neurons. The mean synaptic weight of excitatory projections from the E cells of each local network to the I cells of the other local network is $\frac{c}{N/2}$: $c = 0$ for the BG model network, and $c > 0$ for the FK model network. We model sparse and random connectivity: a small fraction p of the weights are non-zero, and each non-zero weight is independently drawn from a normal distribution with mean $\frac{x}{pN/2}$ and standard deviation $\frac{|x|}{2\,pN}$, where $x = a$ for local excitatory synapses, $x = -b$ for local inhibitory synapses, and $x = c$ for across-network excitatory synapses. We have chosen the standard deviations of the weight distributions to be small enough that we have not observed weights that violate Dale's

Law; if observed, such weights would be set to zero.

We model the dynamics of the neurons with the following linear differential equation:

$$\boldsymbol{T}\frac{d\vec{r}}{dt} = -\vec{r}(t) + \boldsymbol{W}\vec{r}(t) + \vec{I}(t)$$

where $\boldsymbol{T}$ is a diagonal matrix of the time constants of the neurons (normally distributed with mean τ and standard deviation τ/k; again, negative time constants were not observed, but would be set to 1 if observed), $\vec{r}$ is a vector of the activity of the neurons, $\boldsymbol{W}$ is the synaptic weight matrix, and $\vec{I}$ is a vector of the input to the neurons from areas outside LIP. Negative firing rates are not allowed and are rectified to zero (in our simulations, firing rates generally stay positive and do not reach zero). This is a standard phenomenological firing rate model that can be derived as an approximation to biophysically realistic spiking models (Dayan and Abbott, 2005). These dynamics are taken to be modeling trial-averaged firing rates, as we have no knowledge of the trial-to-trial noise correlation among LIP neurons in this task.

The total external input $\vec{I}$ to the neurons at any time is the sum of one or more of the four types of input described in the Results. For each of the four input types, the input to each cell is independently drawn from a uniform distribution, with range of the distribution picked to qualitatively fit the experimentally observed firing rates. The range parameters for fixation input are: (I_{F1}, I_{F2}); transient visual input: (I_{V1}, I_{V2}); sustained visual input: (I_{V1}', I_{V2}'); delay input: (I_{D1}, I_{D2}); expectation input: (I_{E1}, I_{E2}). The transient visual input is modeled to last for 100 ms for the BG model and 40 ms for the FK model. The onset of delay input, as well as the sustained visual input in the FK model, is at the

offset of the transient visual input evoked by a target. The input at any time *t* has two components:

$$\vec{I}(t) = \vec{I}_{determ.}(t) + \vec{I}_{noise}(t)$$

$\vec{I}_{determ.}(t)$ is the deterministic input (the sum of one or more of the four external inputs), and $\vec{I}_{noise}(t)$ is the noise calculated as follows:

$$\vec{I}_{noise}(t) = v\vec{I}_{noise}(t-1) + \vec{I}_{random}(t)$$

v is a parameter with range from 0 to 1, which determines how much the noise is temporally correlated; $\vec{I}_{random}(t)$ is the new noise at time *t*, which is independently drawn at each *t* from a normal distribution with zero mean and standard deviation equal to a fraction *z* times $\vec{I}_{determ.}(t)$.

The inherited surround suppression model is identical to the FK model except in two ways. First, the two local networks are unconnected. Second, whenever one local network receives visual or delay external input, the mean external input to the other local network is reduced by an amount proportional to the mean visual or delay input: the decrease in input to each cell at time *t* is independently picked from a uniform distribution, whose mean is a fraction *u* of the mean visual and/or delay input at time *t* to the activated local network, and whose range is from 0 to twice its mean.

To simulate the experiments, the simulation was run multiple times (41 times for the BG simulation and 27 times for the FK simulation), each time with random

instantiations of connectivity matrices, neuronal time constants, and inputs. One cell is randomly picked from each simulation to form populations the same sizes as the experimental populations.

The model parameters are: $N = 100$, $a = 1.1$, $b = 0.5$, $c = 0.15$, $p = 0.2$, $\tau = 10$, $k = 10/3$, $I_{F1} = 4$, $I_{F2} = 6$, $I_{V1} = 30$ (BG) or 60 (FK), $I_{V2} = 160$ (BG) or 130 (FK), $I_{V1}' = 2$, $I_{V2}' = 4$, $I_{D1} = 5$, $I_{D2} = 65$, $I_{E1} = 2$, $I_{E2} = 10$, $v = 1/30$, $z = 0.97$, $u = 1/30$. The ranges of external inputs were chosen to be consistent with firing rates in the respective top-down and bottom-up areas and to roughly match the simulated LIP firing rates to the data.

Figure 1.1

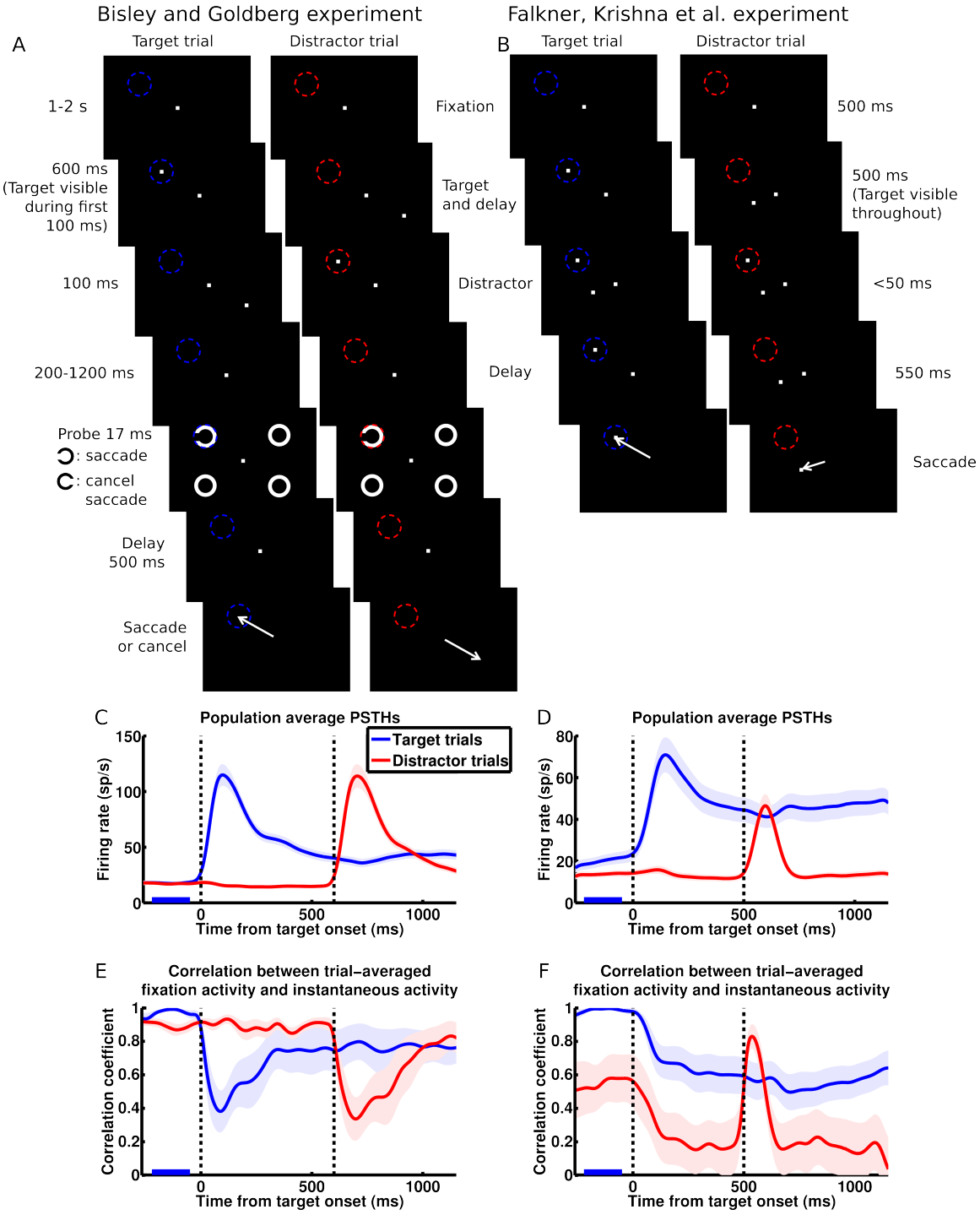

Figure 1.1

The conflicting population dynamics observed by Bisley and Goldberg (2003; BG; left column) and Falkner, Krishna et al. (2010; FK; right column).

(A and B) Task schematics. While the monkey fixates a central spot, a target appears. The monkey is required to hold fixation until the disappearance of the fixation spot, at which time it makes a saccade to the location of the target. During the delay between

target onset and fixation spot disappearance, a task-irrelevant distractor stimulus is flashed. We call a given trial a target trial or distractor trial when the target or distractor, respectively, is in the RF (dashed circles) of the neuron being recorded. In the BG task, the target and distractor are in opposite visual quadrants and equidistant from the fixation spot; in the FK task, either the target or the distractor is in the RF, and the other stimulus is at the location that elicits maximum surround suppression for that RF. In the BG task, between 200 and 1200 ms after the distractor disappears, a probe (a Landolt ring) is flashed at either the target or the distractor location, along with three complete rings elsewhere; a left-facing or right-facing ring instructs the monkey to proceed with or cancel the planned saccade, respectively. In C and E, we only include trials in which the probe appeared at least 700 ms after distractor onset. For task details, see SI section 1.
(C and D) Population average peristimulus time histograms (PSTHs) in the BG (C; n = 41) and FK (D; n = 27) studies. Blue/red traces denote trials in which the target/distractor appears in the RF of the neuron being recorded; every neuron was recorded during both target and distractor trials and contribute to both traces. The first and second vertical dashed lines denote the onset of the target and the distractor, respectively. Shading around traces indicates SEM. PSTHs have been smoothed by convolution with a Gaussian kernel (σ = 30 ms; firing rates and correlations appearing to change before stimuli onset in Fig. 1.1C-F are artifacts of this smoothing).
(E and F) Correlation analysis for the BG (E) and FK (F) datasets. We define a trial-averaged population fixation activity vector $\vec{F}$, each element of which is the activity of one cell on target trials, averaged over trials and over the period from 220 ms to 50 ms before target onset (marked by blue bars in C-F). At each millisecond time point over the course of the target trial (blue traces) or distractor trial (red traces), the correlation coefficient was computed between the trial-averaged population instantaneous activity vector at that point in time and $\vec{F}$. The BG correlation patterns (E; presented in similar format in Ganguli et al., 2008) exhibit one-dimensional dynamics on slow timescales (high correlations during stable fixation activity and delay activity), while the FK correlation patterns (F) markedly deviate from one dimension (on distractor trials, low correlations during stable activity, and transient increase in correlation during distractor visual response). Vertical dashed lines are as in C and D. Shading around traces indicates standard error estimated from 1000 bootstrap samples.
See Fig. 1.S1A for correlations between distractor trials fixation activity and instantaneous activity of the FK data, and Fig. 1.S2A-D for the FK data plotted separately for different reward conditions.

Figure 1.2

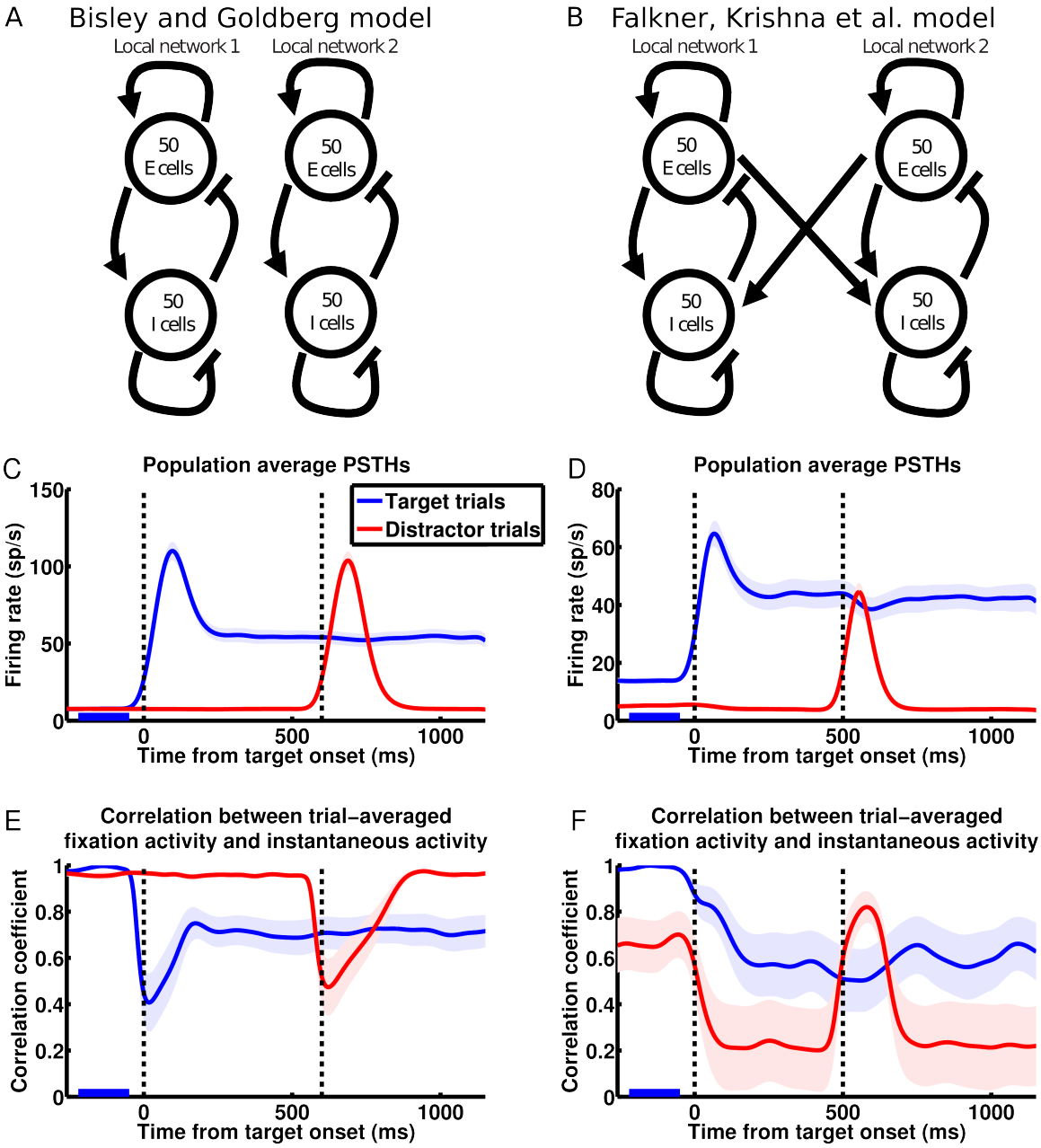

Figure 1.2

Model reproduces the response and network dynamics of Bisley and Goldberg (2003; left column) and Falkner, Krishna et al. (2010; right column).

(A and B) Schematics of the model network connectivity for the BG (A) and the FK (B) scenarios. In both cases we model two recurrently connected local LIP networks corresponding to two RF locations, with each local network consisting of E and I cells. Connectivity within each local network is such that each local network by itself amplifies a single multi-neuronal activity pattern much more strongly than other patterns (see text for details). The FK model network (B) differs from the BG model network (A) in the addition of coupling between the local networks that mediates interaction between the

RFs.
(C and D) Model reproduces LIP activity patterns observed by BG (C; $n = 41$; in all simulation results except where noted, each neuron was "recorded" from a different simulated global network) and FK (D; $n = 27$). Population average PSTHs with same conventions as Fig. 1.1C and D. PSTHs have been smoothed by convolution with a Gaussian kernel ($\sigma = 30$ ms; firing rates and correlations appearing to change before stimuli onset in Fig. 1.2C-F are artifacts of this smoothing).
(E and F) Model reproduces LIP network dynamics observed by BG (E) and FK (F). Correlation analysis with same conventions as Fig. 1.1E and F.
See Fig. 1.S1B for correlations between distractor trials fixation activity and instantaneous activity of the FK simulation, and Fig. 1.S2E-H for separate simulations of the different reward conditions of the FK experiment.

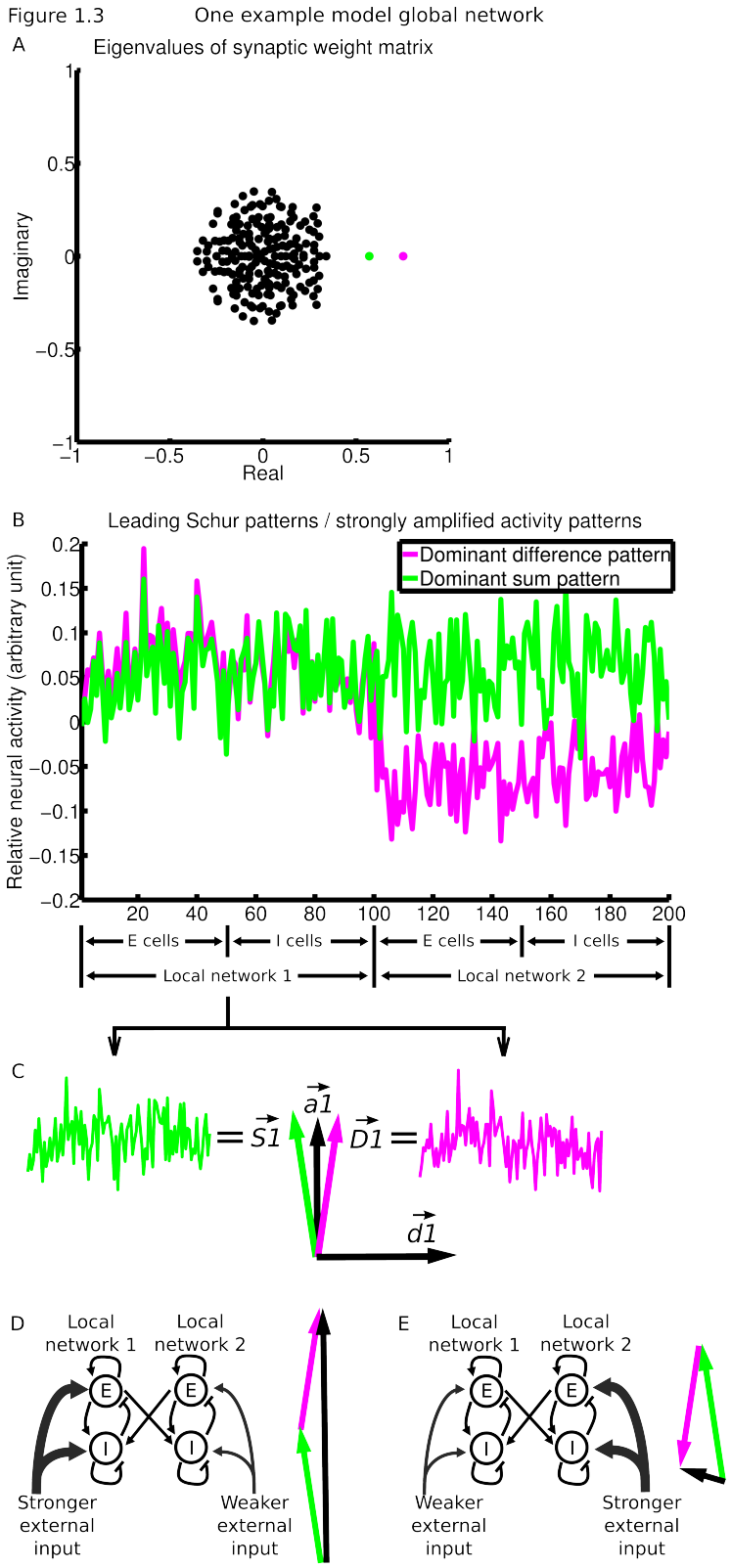

Figure 1.3
Recurrent connectivity strongly amplifies two activity patterns.

(A) The eigenvalue spectrum of the connectivity matrix (matrix plotted in Fig. 1.S3D), for a model network composed of two interconnected local networks of 100 neurons each. Each eigenvalue is associated with a Schur vector, representing a pattern of relative activation across neurons (see text for details). The more positive the real part of an eigenvalue, the more strongly the network amplifies the corresponding Schur activity pattern. Two patterns (magenta and green) are more strongly amplified than others and are plotted in B.
(B) Relative activation across neurons in the dominant difference pattern (differential activation of the two local networks; magenta) and the dominant sum pattern (equal activation of the two local networks; green), or equivalently, the two leading Schur vectors of the connectivity matrix. The difference/sum pattern is driven by the difference/sum of the mean inputs to the two local networks. Note the similarity of the two patterns across cells of the same local network.
(C) The network 1 portion of the sum ($\vec{S}1$) and difference ($\vec{D}1$) patterns can be represented as vectors in the two-dimensional space they define. We can take the axes of the 2D space to be $\vec{a}1$, a vector proportional to the average of $\vec{S}1$ and $\vec{D}1$, and $\vec{d}1$, a vector proportional to their difference.
(D and E) When network 1 receives stronger (D)/weaker (E) mean external input than local network 2, $\vec{S}1$ is activated positively, and $\vec{D}1$ is activated positively (D)/negatively (E). Thus, the $\vec{a}1$ components of $\vec{S}1$ and $\vec{D}1$ add (D)/cancel (E), while the $\vec{d}1$ components of $\vec{S}1$ and $\vec{D}1$ cancel (D)/add (E). The actual activity vectors (black) thus point in very different directions in D and E.
See Fig. 1.S3 for analysis of the feedforward connections between the Schur patterns, comparisons of the directions of dominant activity patterns, and demonstrations of the equivalence of complex sum pattern pairs with single real sum patterns.

Figure 1.4 Simulation of a single global network

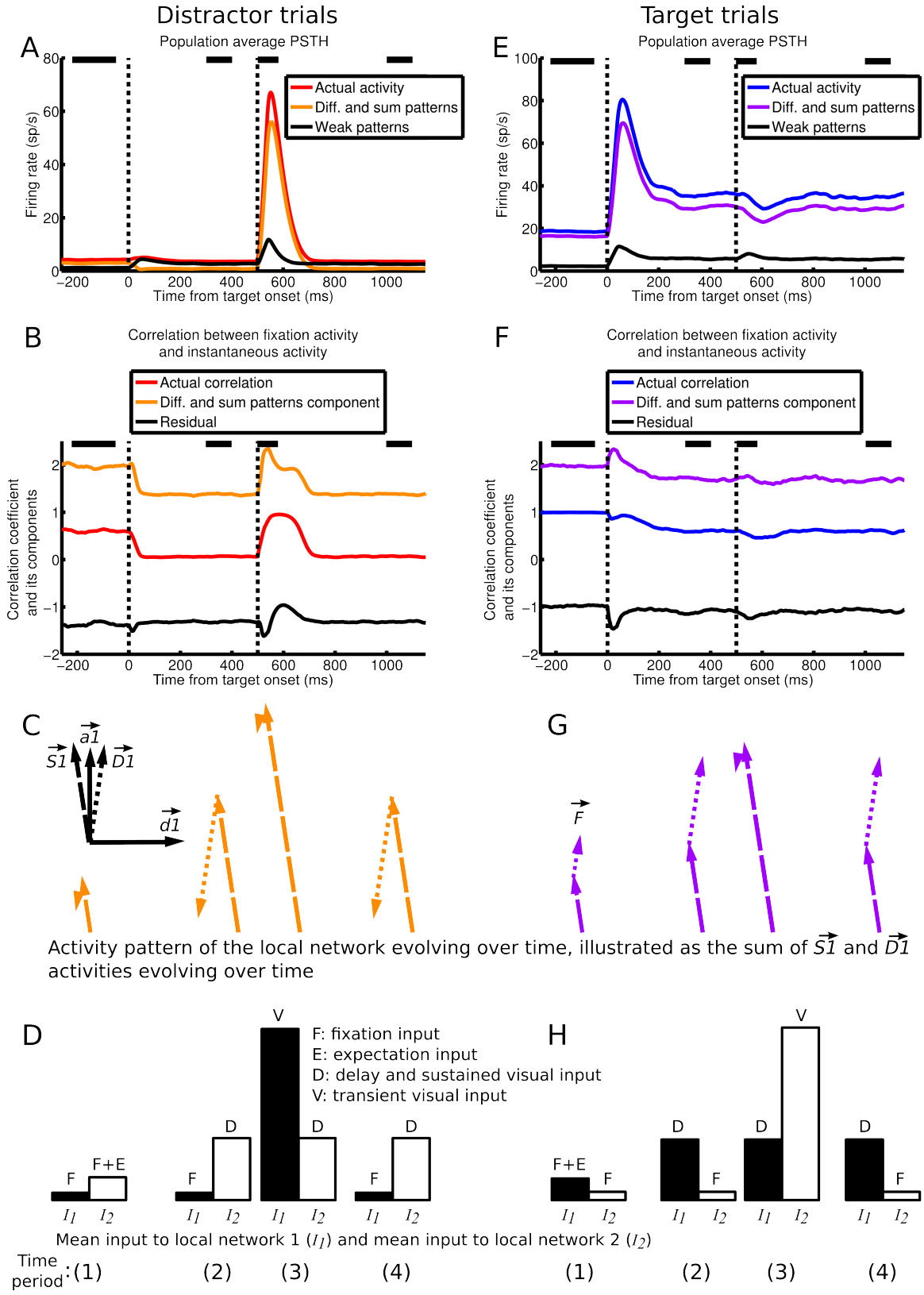

Figure 1.4
Two multi-neuronal activity patterns explain LIP dynamics. One global network composed of local networks 1 and 2 is simulated, and the dynamics in local network 1 on distractor trials (A-D) and target trials (E-H) are analyzed.
(A and E) $\vec{S}1$ and $\vec{D}1$ patterns dominate activity. Population average activity (red/blue), its component in the space of $\vec{S}1$ and $\vec{D}1$ patterns (orange/purple), and its component in the space of all other patterns (black) on distractor/target (A/E) trials (n = 100). In A the orange and black traces add up to the red trace, and in E the purple and black traces add up to the blue trace.
(B and F) Actual correlation (red/blue), $Corr_{sum,diff}$ (orange/purple, the component of correlation due to the $\vec{S}1$ and $\vec{D}1$ patterns alone), and $Corr_{sum,diff}$ and $Corr_{residual}$ (black, the residual component) on distractor/target (B/F) trials. The orange and black traces add up to the red trace, and the purple and black traces add up to the blue trace. $Corr_{sum,diff}$ mirrors the salient ups and downs in the actual correlation, while $Corr_{residual}$ largely does not change with time—thus, the changes in actual correlation over a trial are largely due to the $\vec{S}1$ and $\vec{D}1$ patterns. See Results for how the correlation was broken down into two components. Note that the actual correlation, but not $Corr_{sum,diff}$ or $Corr_{residual}$, is restricted to lie within -1 and 1.
(C top-left inset) The two-dimensional space spanned by the two N-dimensional dominant activity patterns of network 1, $\vec{S}1$ (dashed vector) and $\vec{D}1$ (dotted vector), with $\vec{a}1$ and $\vec{d}1$ as axes.
(C and G) The evolving activation of $\vec{S}1$ (dashed vectors) and $\vec{D}1$ (dotted vectors) activity during distractor (C; orange vectors) and target (G; purple vectors) trials, as a result of the evolving inputs illustrated in D and H. For each trial type, activity in the $\vec{S}1$ and $\vec{D}1$ directions are each averaged over each of four time periods (spanned by black bars in A, B, E, and F), and are illustrated in their two-dimensional subspace, with the relative lengths of and the angle between $\vec{S}1$ and $\vec{D}1$ activity accurately rendered. In this 2D space, at a given time, the activity pattern across cells of network 1 is the vector sum of $\vec{S}1$ and $\vec{D}1$ activity at that time. Thus, $\vec{F}$, the vector of target trial fixation activities, is the vector sum of the $\vec{S}1$ and $\vec{D}1$ activity vectors at time (1) in panel G. The angle between $\vec{F}$ and the vector sum of $\vec{S}1$ and $\vec{D}1$ activity at a given time period generally determines the actual correlation at that time: the larger the angle, the lower the correlation, and vice versa (see Fig. 1.S4 for the precise relationship between the vectors and correlation.). For example, the angle between the vector sum during the delay on distractor trials (vector sum of the $\vec{S}1$ and $\vec{D}1$ activity at time (2) in C) and $\vec{F}$ is large, so the correlation during that time period is low; the vector sum following distractor onset on distractor trials (vector sum of $\vec{S}1$ and $\vec{D}1$ activity at time (3) in C) points in similar directions as $\vec{F}$, so the correlation during that time period is high.
(D and H) The relative input to local networks 1 and 2 during the four time periods on distractor (D) and target (H) trials. Black and white bars denote the mean input to network 1 (I_1) and mean input to local network 2 (I_2), respectively. As illustrated in Fig.

1.3D and E, during each time period, the sum of (difference between) the black and white bars largely determine the magnitude and direction of $\vec{S}1$ ($\vec{D}1$) activity, which are plotted directly above the bars in C and G. Note that the inputs illustrated here predict the steady state activation of $\vec{S}1$ and $\vec{D}1$ *if* the inputs are sustained, which is the case for time periods (1), (2), and (4); however, over time period (3), $\vec{S}1$ and $\vec{D}1$ are not at the steady state predicted by their input, because the input is transient. See text for explanations of why for the same time period on distractor and target trials, the magnitude of $\vec{S}1$ (or $\vec{D}1$) activity might be different, even though I_1 and I_2 are simply flipped in one trial type compared to the other.

Figure 1.5

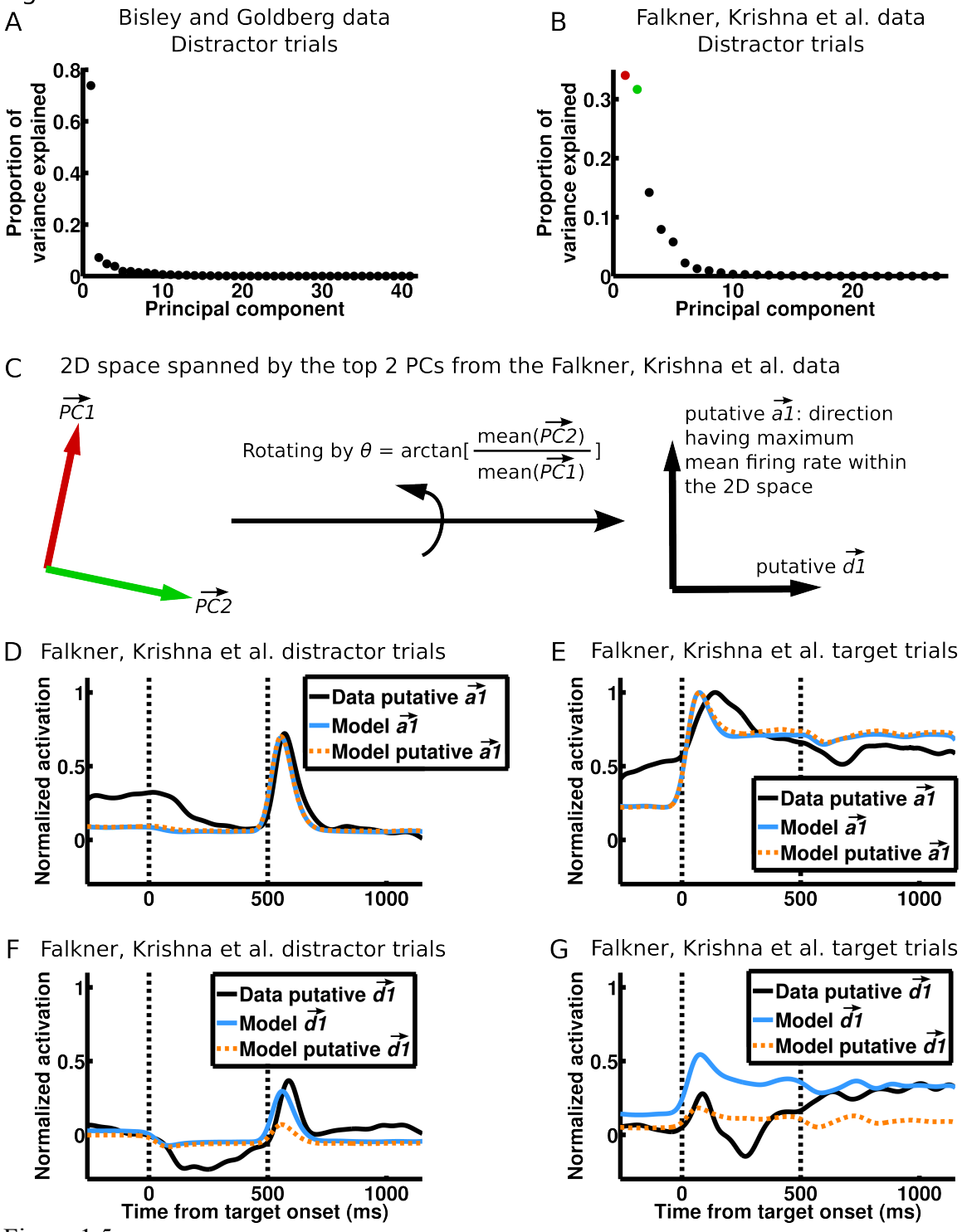

Figure 1.5

Direct evidence for two-dimensional dynamics in the Falkner, Krishna et al. dataset. (A-B) PCA, where the variables are neurons and the observations are the instantaneous activity vectors during distractor trials, for the BG (A) and FK (B) datasets. Activity vectors during the transient visual responses to the distractor (600-1100 ms after target

onset for BG, 450-750 ms after target onset for FK) were not included for this analysis because they involve activation of weak patterns. 74% of variance is explained by one PC in BG, while a comparable proportion is explained by two PCs in FK, consistent with one-dimensional dynamics in BG and two-dimensional dynamics in FK.
(C) We hypothesize that the 2D space spanned by the top 2 PCs (colored as in B) in the FK data is the 2D space of $\vec{S}1$ and $\vec{D}1$ (Fig. 1.3C and Fig. 1.4C, G). We further hypothesize that the direction having the maximum mean firing rate within the 2D space of the 2 PCs is close to the direction of $\vec{a}1$, since $\vec{a}1$ is a direction representing concerted firing of neurons in a local network. We can thus find the putative $\vec{a}1$ and $\vec{d}1$ of the FK data by rotating the two PCs by an angle of arctan[mean($P\vec{C}1$)/mean($P\vec{C}2$)], where mean(•) denotes mean over the elements of a vector.
(D-G) The activation of $\vec{a}1$ (D-E) and $\vec{d}1$ (F-G) on FK distractor trials (D and F) and target trials (E and G). In D-E, the data putative $\vec{a}1$ was derived as in C. To determine activation in the model, one cell was "recorded" from each of multiple simulated global networks to form the model population. To determine the model $\vec{a}1$, suppose the *i*th cell of the model population is the *j*th cell from local network 1 of the *i*th global network. Then the *i*th element of the model $\vec{a}1$ is the *j*th element of the actual $\vec{a}1$ of the *i*th global network. The model putative $\vec{a}1$ was derived as in C but from the model population. The $\vec{d}1$ directions are determined similarly in F-G. Each set of activations (e.g. the four activation traces of data putative $\vec{a}1$ and $\vec{d}1$ on target and distractor trials comprise a set) is normalized by its peak $\vec{a}1$ activation on target trials—thus, D-G share the same scale. Vertical dashed lines denote the onsets of the target and distractor.

Figure 1.6

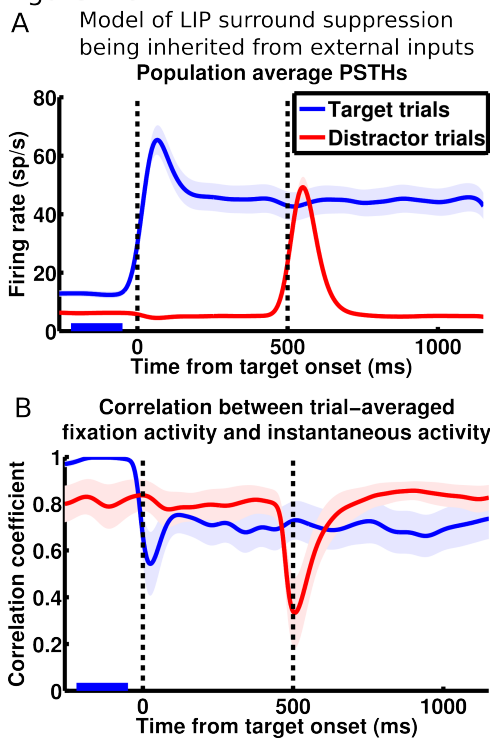

Figure 1.6

Model of inherited surround suppression cannot reproduce observed network dynamics. The model: two LIP local networks are uncoupled; whenever one local network receives visual or delay input, the external input to the other local network is reduced.

(A) Population average PSTHs (n = 27; same conventions as Fig. 1.1D) shows that this model reproduces firing rates observed by FK (Fig. 1.1D) during surround interactions.

(B) This model fails to reproduce network dynamics observed by FK (Fig. 1.1F) during surround interactions. Correlation analysis with same conventions as Fig. 1.1F.

See Fig. 1.S5 for three other models that achieve surround suppression by changing the external input to an isolated local network, all of which fail to reproduce the FK network dynamics.

Figure 1.7
Models with varying levels of surround suppression

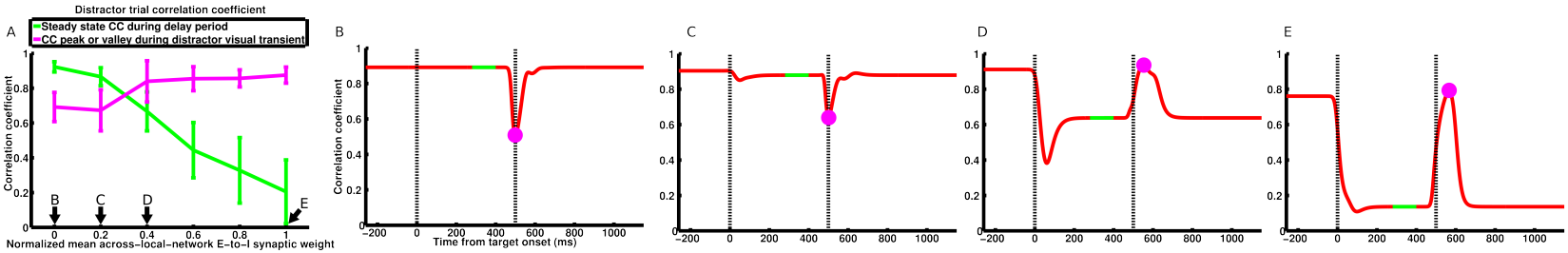

Figure 1.7
Model predictions for the network dynamics underlying different levels of surround suppression.
(A) Two salient features (illustrated in B-E) of distractor trials correlation as functions of the mean across-local-network E-to-I synaptic weight. Normalized mean weights of 0 and 1 are the values used in our BG and FK models, respectively (e.g., Fig. 1.2); intermediate weight values produce intermediate levels of surround suppression (data not shown). The delay period steady-state correlation coefficient is defined as the average correlation from 280 to 400 ms after target onset. The correlation coefficient peak/valley is defined as the maximum correlation from 500 to 600 ms when the correlation is transiently rising, or the minimum correlation from 500 to 600 ms when the correlation is transiently dropping. Error bars are standard deviations across simulations (n = 100 simulations for each value of mean synaptic weight; the parameters of each simulation are independently and randomly drawn). Note that for the mean weight of 0, the correlation coefficient valley plotted here is less deep than that in our BG model (Fig. 1.2E), because all simulations in this figure use the FK visual input parameters (see SI section 6 for the effects of visual input on correlations).
(B-E) Distractor trials correlations from representative simulations of networks with the different levels of coupling indicated by arrows in A. Green traces denote the interval over which the steady-state correlation coefficient in A is calculated, and the magenta dots denote the correlation coefficient peaks/valleys in A. Plotted with same conventions as Fig. 1.1F.
See Fig. 1.S8 for characterization of the dominant activity patterns at different levels of surround suppression.

Figure 1.S1

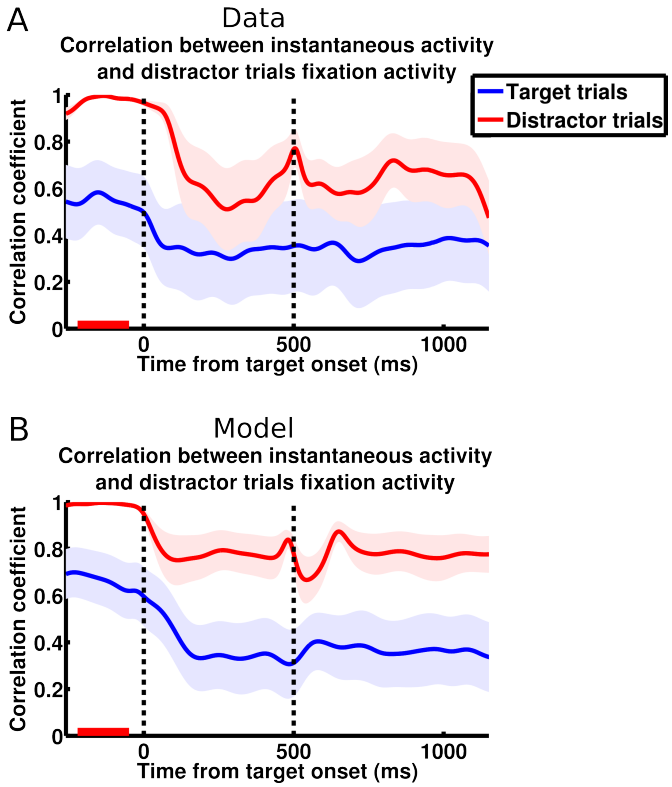

Figure 1.S1
Correlations between distractor trial fixation activity and instantaneous activity of the Falkner, Krishna et al. (FK) data and simulation.
(A) Correlation analysis on the FK dataset, calculated using distractor trial fixation activity. The correlations are calculated similarly to that in Fig. 1.1F, except that fixation activity is averaged over distractor trials (over the period from 220 ms to 50 ms before target onset, marked by the red bar) instead of target trials. Same conventions as Fig. 1.1F.
(B) Correlation analysis on the FK simulation results (same simulated dataset as that in Fig. 1.2D and F), calculated using distractor trial fixation activity. Same conventions as Fig. 1.1F.

Figure 1.S2

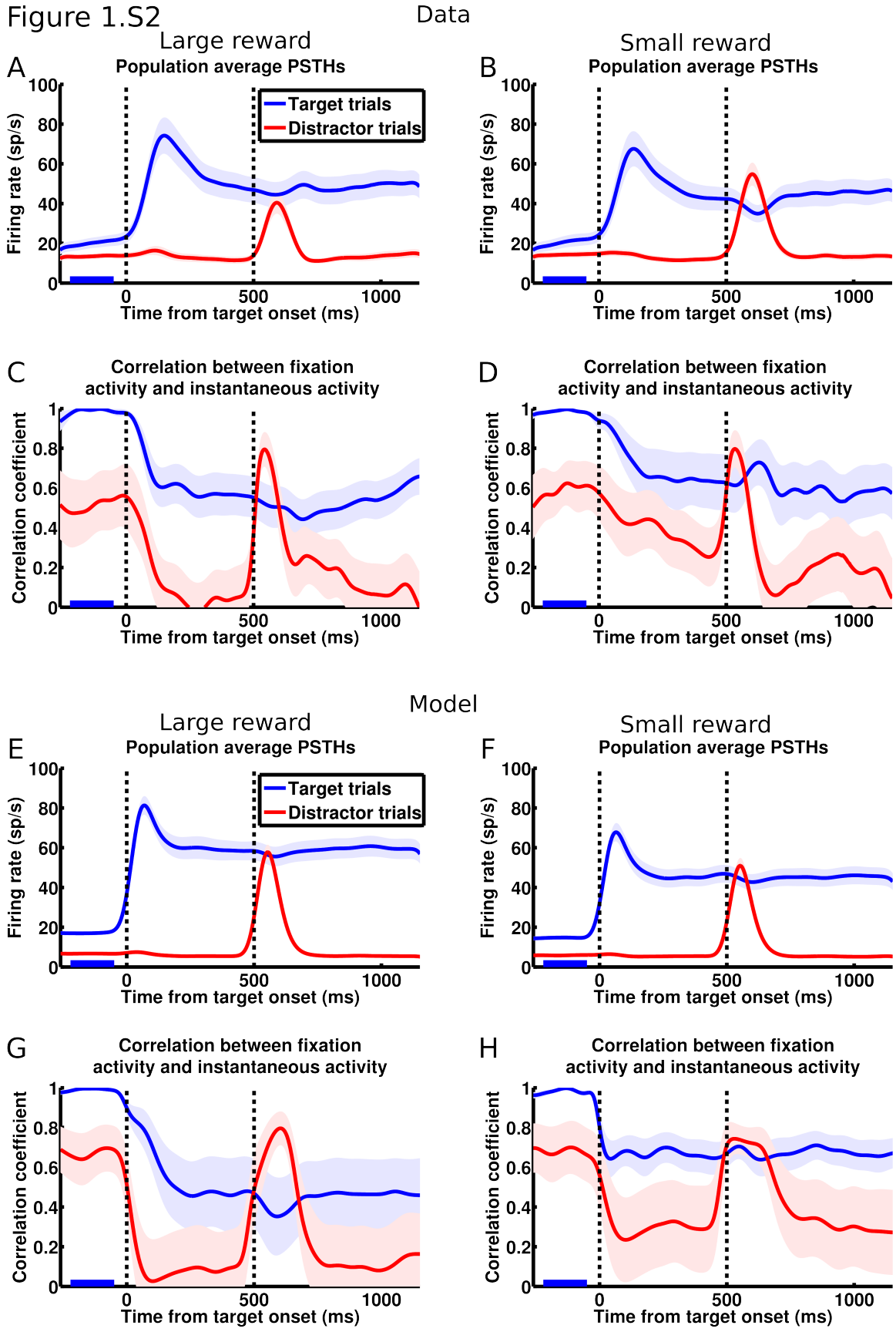
Data
Large reward
A
Population average PSTHs
Target trials
Distractor trials
Firing rate (sp/s)
Time from target onset (ms)
Small reward
B
Population average PSTHs
Firing rate (sp/s)
Time from target onset (ms)
C
Correlation between fixation activity and instantaneous activity
Correlation coefficient
Time from target onset (ms)
D
Correlation between fixation activity and instantaneous activity
Correlation coefficient
Time from target onset (ms)
Model
Large reward
E
Population average PSTHs
Target trials
Distractor trials
Firing rate (sp/s)
Time from target onset (ms)
Small reward
F
Population average PSTHs
Firing rate (sp/s)
Time from target onset (ms)
G
Correlation between fixation activity and instantaneous activity
Correlation coefficient
Time from target onset (ms)
H
Correlation between fixation activity and instantaneous activity
Correlation coefficient
Time from target onset (ms)

Figure 1.S2
The Falkner, Krishna et al. data and simulation results, plotted separately for different reward conditions.
(A and B) Population average PSTHs on large reward (A) and small reward (B) trials (n = 27) in the FK dataset. Same conventions as Fig. 1.1D.
(C and D) Correlation analysis on large reward (C) and small reward (D) trials in the FK dataset. Correlations are calculated similarly to that in Fig. 1.1F, except that fixation activity is averaged over only target trials with large reward (C) or small reward (D). Same conventions as Fig. 1.1F.
(E and F) Activity in separate simulations of the large reward (E) and small reward (F) conditions of the FK experiment (n = 27). Large and small rewards were modeled by using delay input ranges (parameters I_{D1} and I_{D2}) of 7 – 67 and 2 – 62, respectively. Population average PSTHs with same conventions as Fig. 1.1D.
(G and H) Correlations from separate simulations of the large reward (G) and small reward (H) conditions of the FK experiment. Correlations are calculated similarly to that in Fig. 1.1F, except that fixation activity is averaged over only target trials with large reward (G) or small reward (H). Same conventions as Fig. 1.1F.

Figure 1.S3

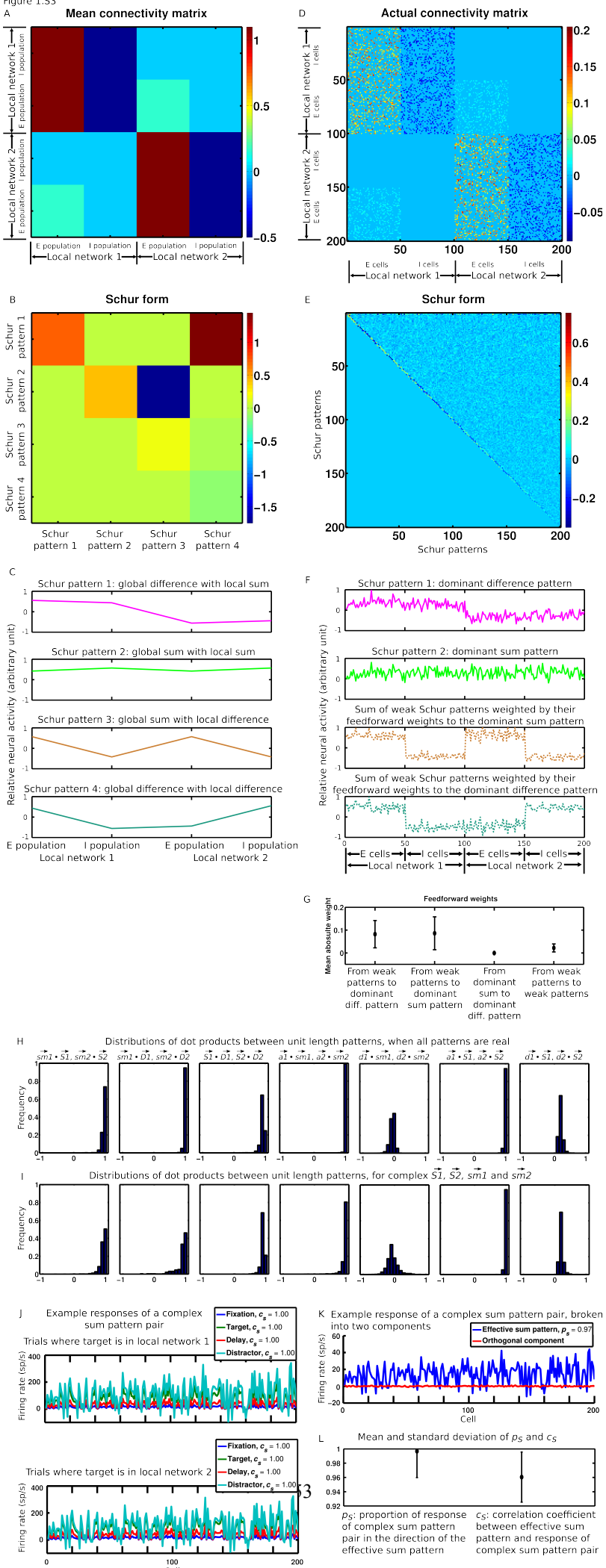
A
Mean connectivity matrix
Local network 1
Local network 2
E population
I population
D
Actual connectivity matrix
E cells
I cells
B
Schur form
Schur pattern 1
Schur pattern 2
Schur pattern 3
Schur pattern 4
E
Schur patterns
C
Schur pattern 1: global difference with local sum
Schur pattern 2: global sum with local sum
Schur pattern 3: global sum with local difference
Schur pattern 4: global difference with local difference
Relative neural activity (arbitrary unit)
F
Schur pattern 1: dominant difference pattern
Schur pattern 2: dominant sum pattern
Sum of weak Schur patterns weighted by their feedforward weights to the dominant sum pattern
Sum of weak Schur patterns weighted by their feedforward weights to the dominant difference pattern
G
Feedforward weights
Mean absolute weight
From weak patterns to dominant diff. pattern
From weak patterns to dominant sum pattern
From dominant sum to dominant diff. pattern
From weak patterns to weak patterns
H
Distributions of dot products between unit length patterns, when all patterns are real
Frequency
I
Distributions of dot products between unit length patterns, for complex $\vec{S1}$, $\vec{S2}$, $\vec{sm1}$ and $\vec{sm2}$
J
Example responses of a complex sum pattern pair
Trials where target is in local network 1
Trials where target is in local network 2
Fixation, c_s = 1.00
Target, c_s = 1.00
Delay, c_s = 1.00
Distractor, c_s = 1.00
Firing rate (sp/s)
Cell
K
Example response of a complex sum pattern pair, broken into two components
Effective sum pattern, p_s = 0.97
Orthogonal component
L
Mean and standard deviation of p_S and c_S
p_S: proportion of response of complex sum pattern pair in the direction of the effective sum pattern
c_S: correlation coefficient between effective sum pattern and response of complex sum pattern pair

Figure 1.S3
Analysis of the Schur form of the connectivity matrix, comparisons of the directions of dominant activity patterns, and demonstrations of the equivalence of complex sum pattern pairs with single real sum patterns. See Supplemental Information sections 2-5 for details.
(A) A mean population connectivity matrix between the E and I populations of two local networks.
(B) The Schur form of the mean population connectivity matrix.
(C) The four Schur patterns of the mean population connectivity matrix.
(D) An actual connectivity matrix. The same one analyzed in Fig. 1.3.
(E) The Schur form of the actual connectivity matrix.
(F) The two leading Schur patterns (the dominant difference and sum patterns in Fig. 1.3), and sums of all other Schur patterns weighted by their feedforward weights to each of the two leading Schur patterns, respectively. The leading Schur patterns and the weighted sums correspond to the Schur patterns of the mean population connectivity matrix (C).
(G) Comparison of mean absolute feedforward weights in E, with standard deviations. The strongest feedforward connections are those from the weak patterns to the leading patterns. The feedforward weight from the dominant sum pattern to the dominant difference pattern is small, making these two patterns effectively independent.
(H) Distributions of dot products over 1000 random instantiations of weight matrices where the vectors in the dot products are real. In the following definitions, x can be 1 or 2 to specify network 1 or 2. $s\vec{m}\,x$: the slow mode of a local network. $\vec{S}\,x$ (or $\vec{D}\,x$): the portion of the global sum (or difference) pattern restricted to cells of a single local network. $\vec{a}\,x$ (or $\vec{d}\,x$): the average (or difference) of $\vec{S}\,x$ and $\vec{D}\,x$. All patterns are normalized to have unit vector length. The overall sign of each $\vec{S}\,x$, $\vec{D}\,x$, and $s\vec{m}\,x$ vector is defined such that the mean of the vector is positive. Parameters of the weight matrices are given in the Experimental Procedures.
(I) Same as H, but over 1000 random instantiations of weight matrices where at least one of $\vec{S}\,1$, $\vec{S}\,2$, $s\vec{m}\,1$, and $s\vec{m}\,2$ is a pair of complex patterns. Only dot products involving at least one of these complex patterns went into the distributions here. For each complex pattern pair, we calculate the effective real pattern as the steady-state response of the complex pair to a uniform input across cells (i.e., a vector of all ones). The effective real patterns are then used to calculate the dot products.
(J) Example responses of a complex sum pattern pair to the eight different inputs in the task. c_s is calculated from each response as the correlation coefficient between it and the effective sum pattern. High firing rates in the target and distractor responses result from hypothetically sustaining the strong visual input to let the responses reach steady state.
(K) Example response of a complex sum pattern pair to fixation input, broken into response in the effective sum pattern and response in the orthogonal direction. p_s is calculated as the proportion of the total response in the direction of the effective sum pattern.
(L) Means and standard deviations of c_s and p_s, calculated from 8000 responses (8 responses for each of 1000 weight matrices) of complex sum pattern pairs.

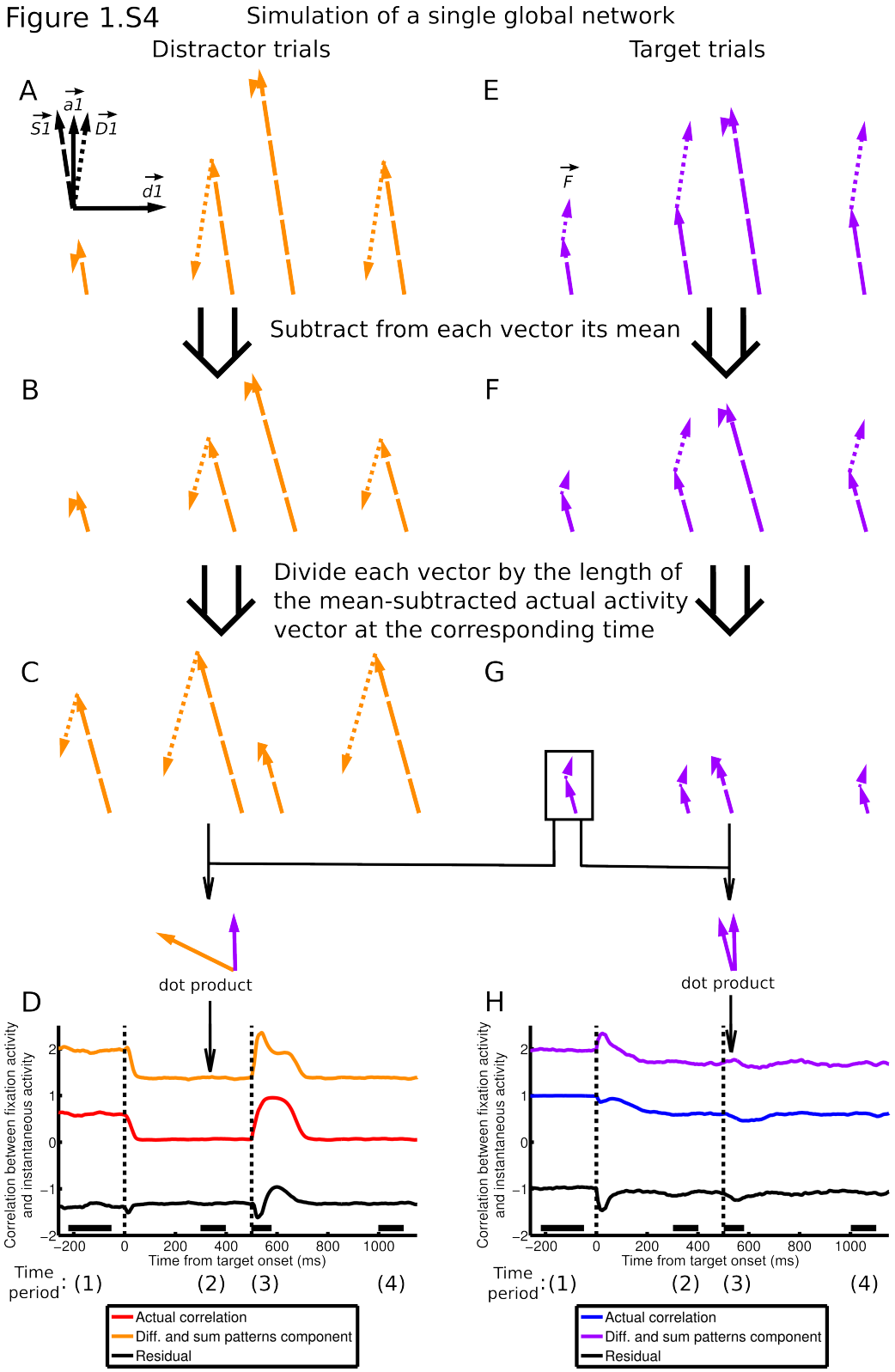

Figure 1.S4

Details of the relationship between $\vec{S}1$ and $\vec{D}1$ and the correlation between fixation and

instantaneous activity during a given time period. A-D are distractor trials; E-H are target trials.

(A inset) The two-dimensional space spanned by the two dominant activity patterns of one local network, $\vec{S}1$ (dashed vector) and $\vec{D}1$ (dotted vector). Replotted from Fig. 1.4C inset.

(A and E) The evolving activation of $\vec{S}1$ and $\vec{D}1$. For each trial type, $\vec{S}1$ and $\vec{D}1$ are each averaged over each of four time periods (spanned by black bars in D and H), and are illustrated in their two-dimensional subspace, where the relative lengths of and the angle between $\vec{S}1$ and $\vec{D}1$ are preserved and accurately rendered. The $\vec{S}1$ and $\vec{D}1$ components of $\vec{F}$, the vector of target trial fixation activities, are labeled in E. Replotted from Fig. 1.4C and G.

(B and F) For each vector in A and E, its mean was subtracted. The resulting mean-subtracted vectors are illustrated in their two-dimensional subspace. Note that the scales of A and E and of B and F are different.

(C and G) Each vector in B and F is normalized by the length of the mean-subtracted actual activity vector at its respective time. Note that B, F, C, and G share the same space and scale. To calculate $Corr_{sum,diff}$ (the $\vec{S}1$ and $\vec{D}1$ component of the correlation coefficient between instantaneous and fixation activity) at a given time period and for a given trial type, first add the two vectors derived from $\vec{S}1$ and $\vec{D}1$ for that time and trial type, and likewise add the two vectors for the fixation period on target trials (boxed in G). Then, $Corr_{sum,diff}$ at that time and on that trial type is the dot product between the two resultant vectors (illustrated for the second time period on distractor trials and the third time period on target trials).

(D and H) Actual correlation (red/blue), $Corr_{sum,diff}$ (orange/purple, the component of correlation due to the $\vec{S}1$ and $\vec{D}1$ patterns alone), and $Corr_{residual}$ (black, the residual component) on distractor/target (D/H) trials. The orange and black traces add up to the red trace, and the purple and black traces add up to the blue trace. See Results for how the correlation was broken down into two components. Replotted from Fig. 1.4B and E. Note that the actual correlation, but not $Corr_{sum,diff}$ or $Corr_{residual}$, is restricted to lie within -1 and 1.

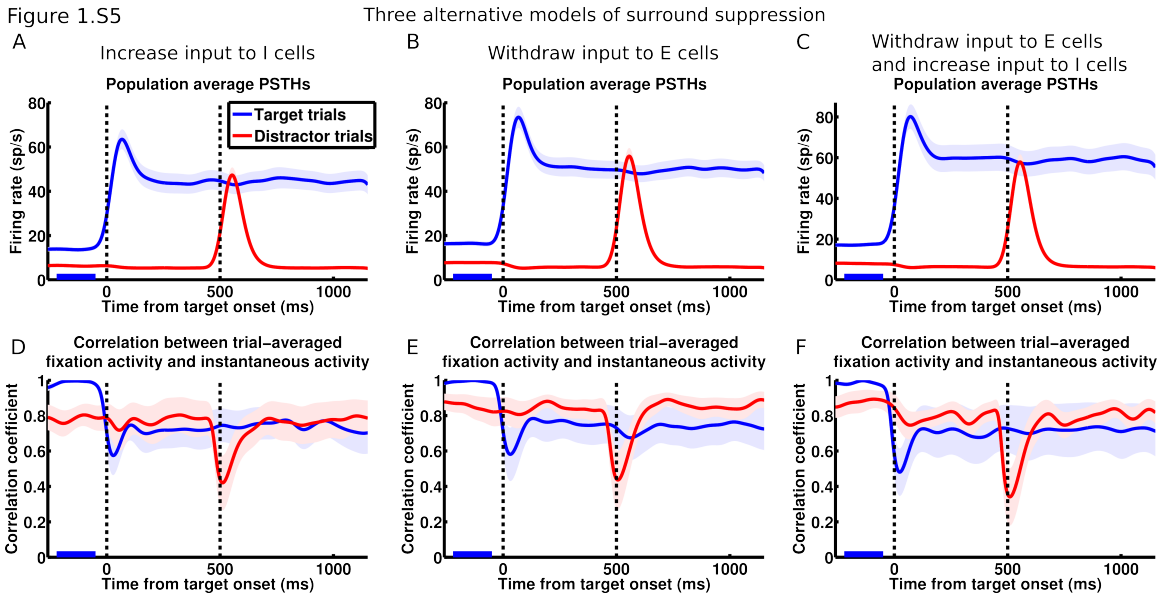

Figure 1.S5

Three models that achieve surround suppression by changing the external input to a local network in LIP that is not coupled to other LIP local networks. Left column: addition of input to I cells alone; middle column: withdrawal of input from E cells alone; right column: addition of input to I cells combined with withdrawal of input from E cells (see Supplemental Information section 7 for details of the models). None reproduces the correlations patterns observed by FK (Fig. 1.1F).

(A-C) Population average PSTHs (n = 27 in each panel; same conventions as Fig. 1.1D) resemble firing rates observed by FK (Fig. 1.1D) during surround interactions.

(D-F) Network dynamics are unlike those observed by FK (Fig. 1.1F) during surround interactions. Correlation analysis with same conventions as Fig. 1.1F.

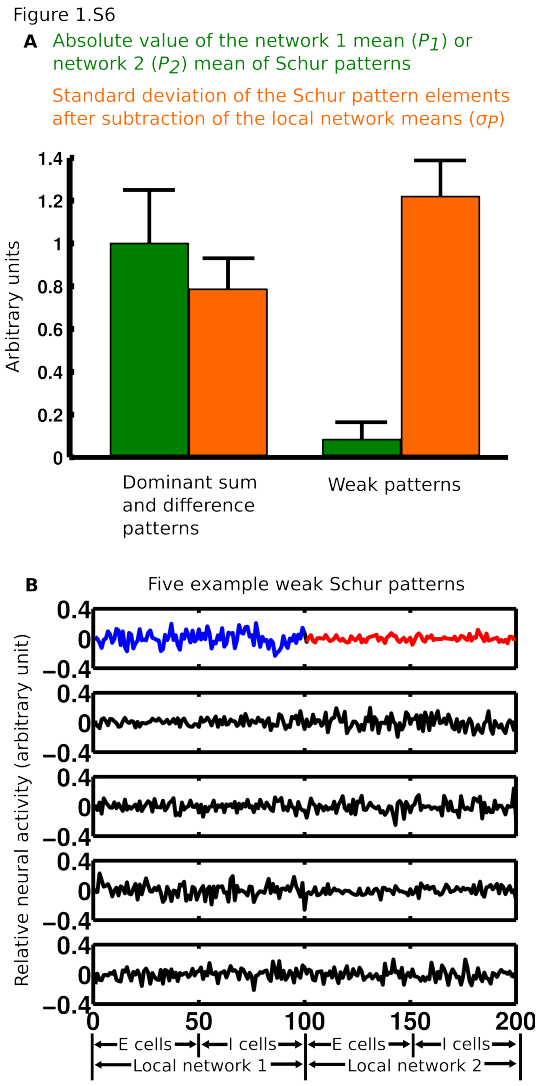

Figure 1.S6

Sum and difference patterns, but not weak patterns, are driven strongly by the mean input to a local network.

(A) 100 global networks were generated. For each Schur pattern of each global network, we examined its two local network portions: the elements corresponding to network 1 (whose mean is P_1) and the elements corresponding to network 2 (whose mean is P_2). For each portion we calculated the absolute value of its mean. Green bar graphs are the mean and standard deviation of all such absolute values for all the dominant patterns, and for all the weak patterns. For example, the absolute values of the means over the blue portion and the red portion of the first example Schur pattern in B are two numbers that went into the green bar for weak patterns plotted here. The low absolute values for the weak patterns compared to the dominant patterns mean that the weak patterns are not strongly driven by the mean input to a local network. For each Schur pattern of each global network, we also subtracted P_1 from its network 1 elements and P_2 from its network 2 elements, and calculated the standard deviation of these mean-subtracted elements (σ_P). Orange bar graphs are the mean and standard deviation of all such σ_P for all the dominant

patterns, and for all the weak patterns. The weak patterns have large σ_P relative to the absolute value of the means, thus they are much more strongly driven by the random input patterns across neurons than by the mean input to a local network.
(B) Five example weak patterns. Note that they represent "random" activation of the neurons (i.e. some neurons increase firing and others decrease firing), unlike the sum and difference patterns (see Fig. 1.3B for examples) which represent concerted activation of all neurons of the same local network (i.e. either all increases firing or all decreases firing).

Figure 1.S7

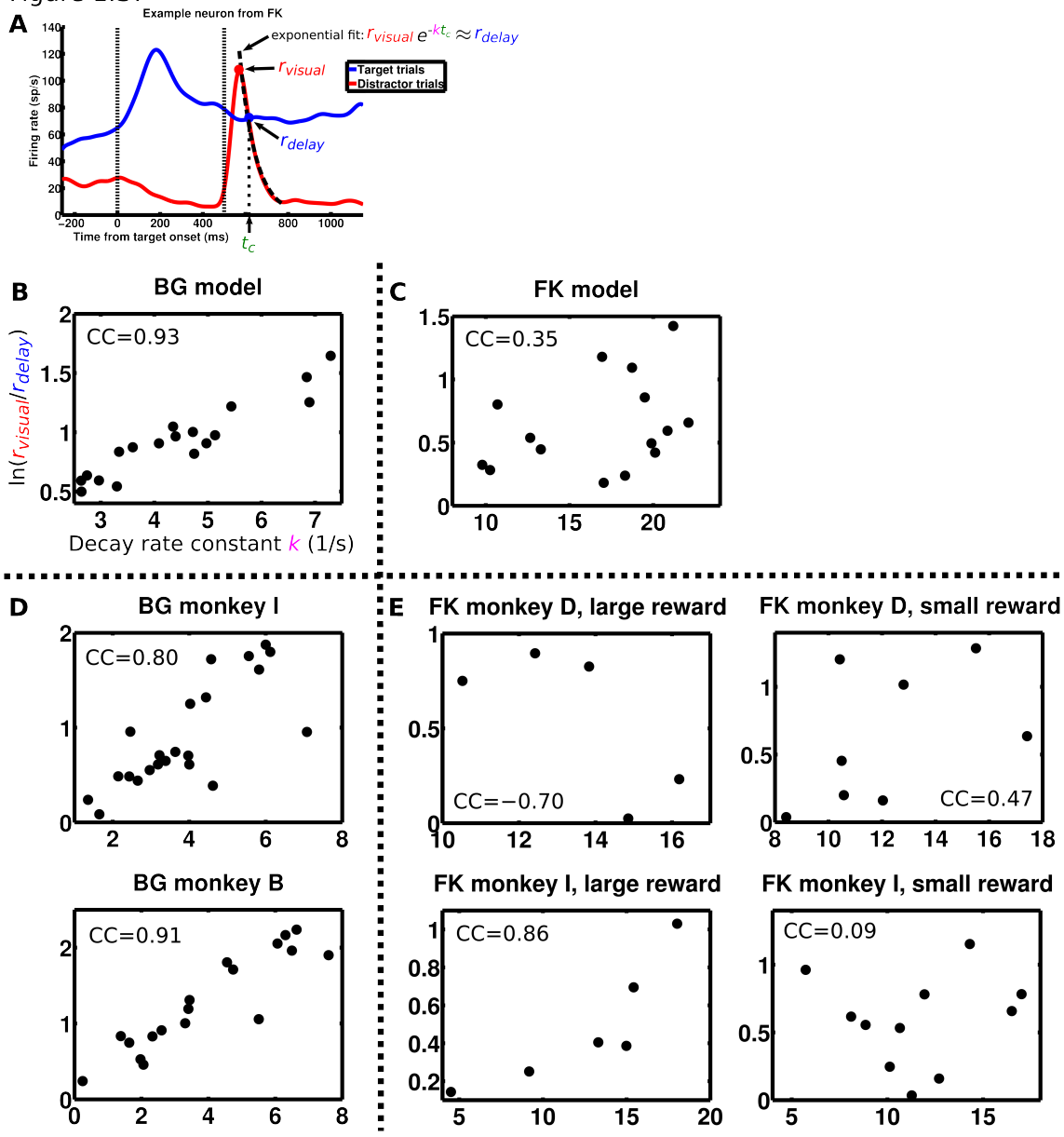

Figure 1.S7

The crossing dynamics of single neurons. This analysis follows Bisley and Goldberg (2006) and Ganguli et al. (2008).

(A) The quantities relevant to the crossing dynamics, illustrated for one example neuron. The decay of the distractor visual response is fit with an exponential function: the peak visual response, r_{visual}, decays exponentially with time constant k, and crosses the delay activity, r_{delay}, at the crossing time, t_c. Single neuron PSTHs plotted with the same conventions as Fig. 1.1D.

(B-E) $ln\,(\,r_{visual}\,/\,r_{delay}\,)$ is plotted against k for BG and FK model and data. Each dot is a single neuron, where the plotted quantities are measured as illustrated in A. Rearranging the equation in A gives $ln\,(\,r_{visual}\,/\,r_{delay}\,) \approx t_c\, k$; thus, the slope of the line connecting each dot to the origin is t_c, the crossing point of that neuron. When $ln\,(\,r_{visual}\,/\,r_{delay}\,)$ and k are

highly correlated as in the BG model and data (B and D), the slopes are similar, meaning that single neurons have similar crossing time. ln (r_{visual} / r_{delay}) and k are less correlated in the FK model and data (C and E), indicating that single neuron crossing times are more variable. D is replotted from Fig. 1E-F of Ganguli et al. (2008). One of the FK monkeys has too few cells (1 cell for large reward, 3 cells for small reward) and is not included in E.

Figure 1.S8

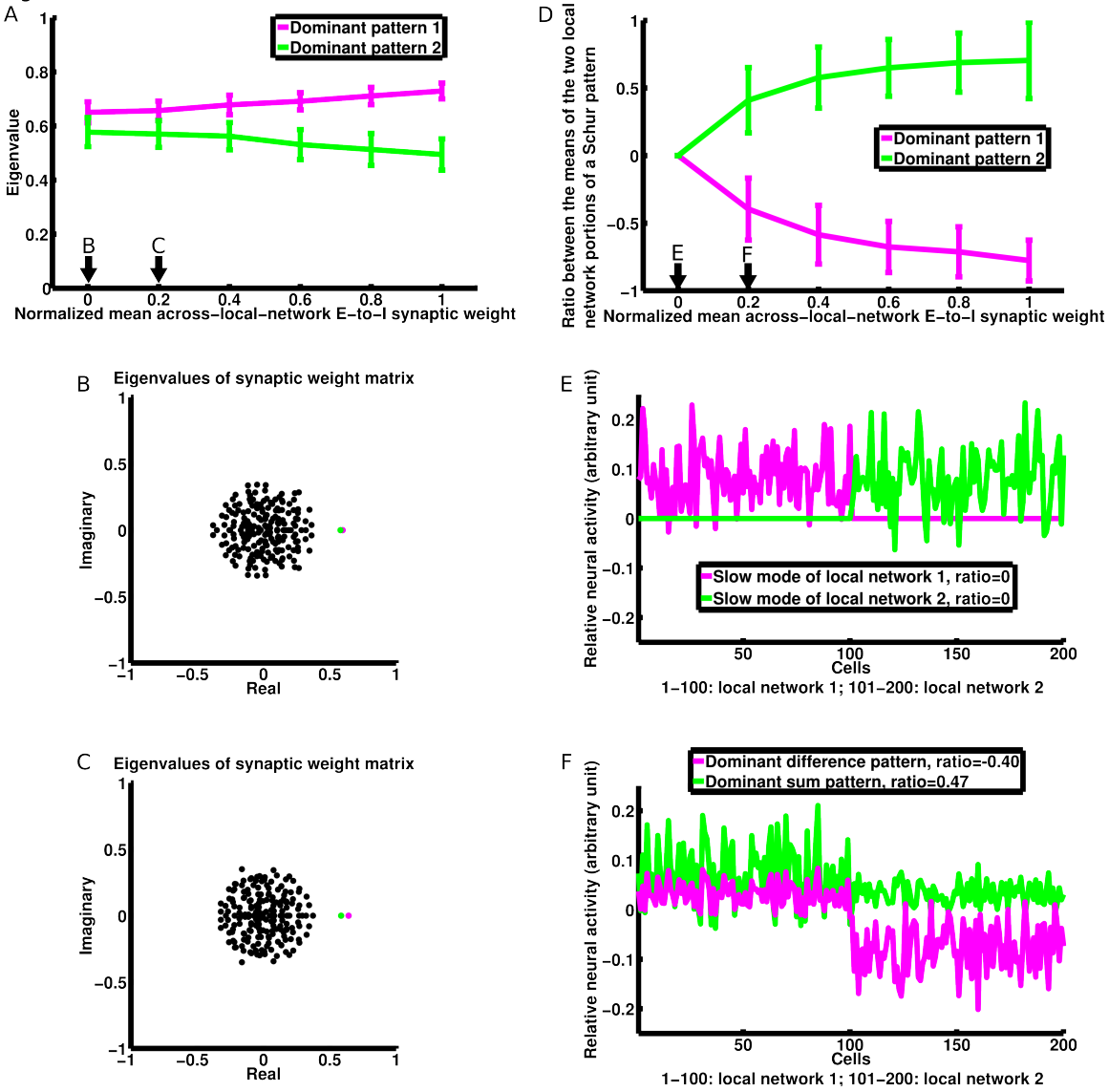

Figure 1.S8

Independent slow modes gradually morph into sum and difference patterns as coupling between local networks strengthens.

(A) The two leading eigenvalues of the global network as functions of the across-local-network E-to-I synaptic weights. As coupling strengthens, one eigenvalue (that of the difference pattern) increases while the other eigenvalue (that of the sum pattern) decreases. Error bars are standard deviations across simulations (n = 100 global networks for each value of mean synaptic weight). The normalized mean weights of 0 and 1 are used in our BG and FK models, respectively (e.g. Fig. 1.2). Weights in-between produce intermediate levels of surround suppression (data not shown). Equations (1) and (2) in Supplemental Information section 3 shows analytically the dependence of the two eigenvalues on the across-local-network weight for the mean population connectivity matrix, which agrees with the eigenvalues of actual connectivity matrices plotted here. Note that with a mean weight of zero, the difference between the two eigenvalues reflect

stochastic differences between the connectivity of the local networks, instead of deterministic differences between sum and difference patterns, as is the case with nonzero mean weights.
(B-C) Representative eigenvalue spectra of networks with the different levels of coupling indicated by arrows in A.
(D) For each dominant pattern (which has 200 elements), we first calculated e_1, the mean over its local network 1 portion (elements 1-100), and e_2, the mean over its local network 2 portion (elements 101-200). Then we calculated the ratio between the e_1 and e_2, with the one that has the larger absolute value in the denominator, so that the ratio ranges between -1 and 1. The ratio for each of the two dominant patterns are plotted as a function of the across-local-network E-to-I synaptic weights. A ratio of 0 indicates that the pattern represents activation of one local network independent of the other local network, i.e. the slow mode of BG. A positive/negative ratio indicates common/differential activation of the two local networks, i.e. the sum/difference pattern. As coupling between the local networks strengthens, the two slow modes morph into the sum and difference patterns. Error bars are standard deviations across simulations (n = 100 global networks for each value of mean synaptic weight).
(E-F) Representative dominant patterns of networks with the different levels of coupling indicated by arrows in D.

1.5 Supplemental Information

Section 1: Task details

At the beginning of each recording session, before task performance, both Bisley and Goldberg (BG) and Falkner, Krishna et al. (FK) isolate an LIP neuron and map out its receptive field (RF). In addition, FK map out a location in the visual field where a stimulus evokes maximum suppression.

In both studies, a monkey initiates a trial by fixating a central spot. After some time (BG: variable between 1 s and 2 s; FK: 500 ms) the saccade target appears. The target disappears after 100 ms in the BG task version, and it stays on in the FK version. After a delay (BG: 600 ms from target onset; FK: 500 ms from target onset), a task-irrelevant distractor stimulus is flashed (duration: BG, 100 ms; FK, <50 ms). After another delay (BG: variable between 700 ms and 1700 ms; FK: 550 ms), the fixation point disappears, and the monkey saccades to the target location for a reward.

In the BG version of the task, the target and the distractor, one of which is in the RF of the neuron being recorded, are in opposite visual quadrants and equidistant from the fixation point (i.e. they are at equal radii from the fixation point, and one is at a location rotated 180 degrees from the other's location). In the FK version of the task, either the target or the distractor is in the RF, and the other stimulus is at the location previously determined to elicit maximum surround suppression. On a given trial, either the target or the distractor is in the RF of the neuron being recorded. In the BG task, the two different trial types are randomly interleaved; in the FK task, the two types of trials were run in blocks.

In the BG version of the task, during the delay period between distractor

presentation and fixation point disappearance, a Landolt ring (a ring with a small segment missing) and three complete rings are flashed simultaneously for 17 ms. These four stimuli are at the target and distractor locations and at the locations rotated 90 degrees about the fixation point from those two locations, so that one is in each of the visual quadrants and all are equidistant from the fixation point. The Landolt ring appeared at either the target or the distractor location. The monkey is required to detect the orientation of the Landolt ring: if the gap is on the right, the monkey needs to cancel the planned saccade and maintain fixation after the fixation point disappears; if the gap is on the left, the monkey can proceed with the planned saccade after the fixation point disappears. The rings were shown at high contrast during neural recordings; they were shown at varying contrasts in separate psychophysical experiments to map contrast thresholds and thus the allocation of attention. In this paper, we only analyzed trials in which the rings appeared more than 700 ms after distractor onset.

In the FK version of the task, in each trial the target is one of two colors, indicating that the reward amount for that trial would be large or small.

Section 2: Analysis of feedforward connections in the Schur form of the connectivity matrix

One common way to examine the influence of the connectivity of a network on its dynamics is through determining the eigenvectors and eigenvalues of the connectivity matrix. The eigenvectors are a set of activity patterns that each excite or inhibit itself but not any of the other patterns. Thus, in a linear model these patterns evolve independently: each evolves according to its own self-connection, independent of the other patterns. The

strength of self-connection of each eigenvector is given by the real part of its corresponding eigenvalue, and so one may expect the eigenvectors whose eigenvalues have the largest real part to dominate the activity of the network.

However, for biological connection matrices composed of separate excitatory and inhibitory neurons, the eigenvectors are not orthogonal (Murphy and Miller, 2009), meaning for example that two eigenvectors with large amplitude can cancel, resulting in small overall activity. These cancellations and related effects can make it difficult to understand neural activities from the independent dynamics of the eigenvectors. Instead, it can be more illuminating to analyze the Schur patterns: an ordered set of *orthogonal* activity patterns derived by orthogonalizing the eigenvectors (Murphy and Miller, 2009; Goldman, 2009). For a given connectivity matrix, there are different sets of Schur patterns, obtained by orthogonalizing the eigenvectors in different orders. For our purpose of finding the dominant activity patterns, we choose the set of Schur patterns that are ordered by their strength of self-connections, from the most self-excitatory to the most self-inhibitory. The self-connections are examined in the main text; here we examine the rest of the connections between the Schur patterns, a set of purely feedforward connections.

To understand the structure of feedforward connections in our connectivity matrices, we first examine the mean population connectivity matrix. This is a 4-by-4 matrix, whose rows and columns denote the excitatory (E) and inhibitory (I) populations of the two local networks, and whose elements are the mean connection strengths between them multiplied by $N/2$ (the number of E or I neurons in each local network). Fig. 1.S3A plots an example mean population connectivity matrix. The four

rows/columns denote: the E population of local network 1, the I population of local network 1, the E population of local network 2, and the I population of local network 2. Each row shows the input weight to the given population from each of the four populations, while each column shows the projection weight from the given population to each of the four populations. Fig. 1.S3B plots the Schur form of this matrix, which shows the connections between the Schur activity patterns or basis vectors (each representing a pattern of activity across the four populations). It shows that in addition to self-connections (non-zero entries on the diagonal, which are the eigenvalues associated with the patterns), there are feedforward connections from activity pattern 3 to pattern 2, and from pattern 4 to pattern 1 (non-zero entries on the upper triangle). What are these activity patterns? Fig. 1.S3C plots the Schur basis vectors. To describe these we will introduce the following terminology. By global sum or difference we mean that the activity patterns of the two local networks are the same or opposite, respectively. By local sum or difference we mean that the activities of the E and I populations within a local network are the same or opposite, respectively. We can see that patterns 1 to 4 represent: global difference with local sum, global sum with local sum, global sum with local difference, and global difference with local difference. The connections from pattern 3 to pattern 2 and from pattern 4 to pattern 1 thus represent local difference patterns feeding into local sum patterns, a manifestation of balanced amplification, which we investigated in Murphy and Miller (2009).

Thus, the dominant activity patterns of the mean population connectivity matrix are patterns 1 and 2, which corresponding to the global sum and difference patterns discussed in the main text, because they are amplified both by strong self-excitation and

by receiving feedforward excitation. Does this structure also hold for the actual connectivity matrix, in which each population consists of many neurons, with weights between members of two populations chosen stochastically? We analyze one actual connectivity matrix, the one examined in Fig. 1.3 of the main text. In Fig. 1.S3D-E, we plot the actual connectivity matrix and its real-valued Schur form. As we have seen in the main paper, the two most strongly self-excitatory patterns of the actual connectivity matrix (plotted in Fig. 1.3B of the main paper and again in Fig. 1.S3F) are still the patterns of global difference with local sum and global sum with local sum, as predicted by the mean population connectivity matrix. We will refer to them here as the dominant difference and dominant sum patterns. The two weaker patterns of the mean population connectivity matrix—the patterns of global sum with local difference and global difference with local difference—are dispersed in the many weakly self-excitatory patterns that are a manifestation of the sparse and random connectivity of the actual connectivity matrix; the feedforward structure of these patterns to the two dominant patterns are hidden, but unchanged. We can reveal the feedforward structure to the dominant difference or sum pattern by summing the less self-excitatory Schur basis vectors (that is, all of the patterns except the dominant difference and sum patterns), each weighted by its feedforward weight to the dominant difference or sum pattern, respectively. The resulting weighted sums are a pattern of global difference with local difference, which feeds into the dominant difference pattern, and a pattern of global sum with local difference, which feed into the dominant sum pattern, just as predicted by the mean population connectivity matrix (Fig. 1.S3F). Furthermore, a comparison of the magnitudes of feedforward weights show that the only strong feedforward connections

are those from the less self-excitatory patterns to the two dominant patterns; in particular, the feedforward connections from the dominant sum pattern to the dominant difference pattern is very weak, making these two dominant patterns essentially independent (Fig. 1.S3G). Thus even before observing network dynamics during simulations (e.g., Fig. 1.4 in main paper), based on the structure of the weight matrix we can predict that the difference and sum patterns would dominate the dynamics of the network.

Section 3: The eigenvalues of the sum and difference patterns

Here we calculate the eigenvalues of the mean population connectivity matrix examined in the last section (e.g. Fig. 1.S3A). This matrix is

$$\begin{pmatrix} a & -b & 0 & 0 \\ a & -b & c & 0 \\ 0 & 0 & a & -b \\ c & 0 & a & -b \end{pmatrix}$$

a is the E weight and -*b* the I weight within a local network, and *c* is the weight of the across-local-network E-to-I connections that mediate surround suppression, and *a*, *b*, and *c* are all positive. The eigenvalues of this matrix are, from the most positive to the most negative,

$$\lambda_D = \frac{1}{2}(a-b+\sqrt{(a-b)^2+4bc})\,, \qquad (1)$$

$$\lambda_S = \frac{1}{2}(a-b+\sqrt{(a-b)^2-4bc})\,, \qquad (2)$$

$$\lambda_3 = \frac{1}{2}(a-b-\sqrt{(a-b)^2-4bc})\,,$$

$$\lambda_4 = \frac{1}{2}(a-b-\sqrt{(a-b)^2+4bc})$$

Each local network by itself has a slow mode when its recurrent excitation dominates recurrent inhibition (i.e. $a > b$). When the two local networks are uncoupled (i.e. $c = 0$, the BG case), λ_D and λ_S are equal and are the slow mode eigenvalues of the independent local networks, while λ_3 and λ_4 are zero. The weak suppressive coupling between the two local networks in the FK case (small, positive c) perturbs these eigenvalues. λ_D and λ_S remain large and positive, and become the eigenvalues of the difference and sum patterns, respectively, while λ_3 and λ_4 remain close to zero. Because the local networks mutually suppress each other, $\lambda_D > \lambda_S$, i.e. the difference pattern is more strongly amplified than the sum pattern by the connectivity. The eigenvalues of the difference and sum patterns of the actual connectivity matrix (e.g. Fig. 1.S3D) are close to λ_D and λ_S respectively (see the Supplementary Materials of Ganguli et al.), while the two weaker patterns associated with λ_3 and λ_4 are dispersed among the many weak patterns of the actual connectivity matrix (Fig. 1.S3F).

If we model the mean population connectivity matrix with more parameters (e.g., separate weight parameters for the E-to-E, E-to-I, I-to-E, and I-to-I connections within a local network, and additional across-local-network E-to-E connections), our formulas for the eigenvalues would become much more complex, but the simple intuition presented above do not change. With parameters of within-local-network weights that result in the isolated local network having a slow mode, the global network would have two and only two dominant patterns. The addition of weak across-local-network mean weights, which are consistent with the weak suppression observed by FK and with the fact that cortical connection density decrease with distance (Markov et al., 2011), acts as a small perturbation, and the sum and difference patterns remain the only two dominant patterns.

Section 4: Directions of sum and difference patterns and the slow mode

In the following we examine the directions of the dominant patterns, using network 1 as an example—but note that networks 1 and 2 are equivalent, and the analysis in Fig. 1.S3H includes both networks). Let $s\vec{m}1$ be the slow mode of network 1 in isolation, and $\vec{S}1$ and $\vec{D}1$ represent the network 1 portions of the sum and difference patterns, respectively. We assign an overall sign to each pattern such that most elements of that pattern are positive, and take all patterns to be normalized to unit vector length. Fig. 1.S3H illustrates the dot products between $s\vec{m}1$ and $\vec{S}1$, and between $s\vec{m}1$ and $\vec{D}1$. $\vec{S}1$ and $\vec{D}1$ are almost but not exactly the same as the slow mode, because coupling between two local networks effectively changes the connectivity underlying the slow modes.

$\vec{S}1$ and $\vec{D}1$ are also similar but not exactly equal to one another (Fig. 1.S3H). This means that, within each local network, the space of strongly amplified activity patterns—the space of activity patterns composed of all weighted sums of $\vec{S}1$ and $\vec{D}1$ —is two-dimensional. Within a local network, a convenient orthogonal pair of vectors to serve as a basis for this two-dimensional space is a vector $\vec{a}1$ proportional to the average of $\vec{S}1$ and $\vec{D}1$, $\vec{a}1 \propto \vec{S}1 + \vec{D}1$, and a vector $\vec{d}1$ proportional to their difference, $\vec{d}1 \propto \vec{S}1 - \vec{D}1$ (Fig. 1.3C). Again, we take $\vec{a}1$ and $\vec{d}1$ to be of unit length. The average of $\vec{S}1$ and $\vec{D}1$ is almost precisely the slow mode, and the difference between $\vec{S}1$ and $\vec{D}1$

is almost orthogonal to the slow mode (Fig. 1.S3H). The vector $\vec{d}\,1$ is the second strongly amplified direction of the activity of a local network, in addition to $\vec{a}\,1$ (which is essentially the slow mode). When $\vec{S}\,1$ is activated, activity is driven strongly in the $\vec{a}\,1$ direction (dot product between $\vec{a}\,1$ and $\vec{S}\,1$ tends to be high, see Fig. 1.S3H), i.e. the direction of the slow mode, and less strongly in the $\vec{d}\,1$ direction (dot product between $\vec{a}\,1$ and $\vec{D}\,1$ tends to be low, see Fig. 1.S3H). It is similar when $\vec{D}\,1$ is activated, which has an identical $\vec{a}\,1$ component and an equal and opposite $\vec{d}\,1$ component as $\vec{S}\,1$.

Section 5: Equivalence of complex sum pattern pairs with single real sum patterns

With the connectivity parameters in the main text, in a small proportion of random instantiations of connectivity matrices, two complex patterns (which are complex conjugates in the eigenvector basis) take the place of the single real global sum pattern. When recurrent excitation is sufficiently stronger than inhibition, all random instantiations of connectivity matrices have real sum patterns, and when excitation is weaker (while still being stronger than inhibition, ensuring the existence of slow modes), complex sum pattern pairs are more frequent. Similarly, the slow mode of an isolated local network can also be a complex pattern pair.

A complex conjugate pair introduces two slowly-decaying patterns of neural activation in place of the single pattern corresponding to a real sum pattern or a real slow mode. However, our analysis remains unchanged, because activation of a complex conjugate pair in response to our various input patterns is very largely confined to a

single dimension, which we call the effective sum pattern or the effective slow mode. We define the effective sum pattern for a complex sum pattern pair (or effective slow mode for a complex slow mode pair) to be the steady-state response of the complex pattern pair to a uniform input across cells of the network (i.e., a vector of $2N$ ones for the sum pattern pair or N ones for the slow mode pair), normalized to a vector length of one. The near complete overlap of dot product distributions calculated with real patterns and effective patterns (Fig. 1.S3H and I) shows that the effective patterns would behave the same way as the real patterns analyzed in the main text.

In response to inputs used to simulate the experiments, the response of complex sum pattern pairs or complex slow mode pairs corresponds almost perfectly to their effective sum patterns. To illustrate this, we simulated 8000 such responses for complex sum patterns (for each of 1000 weight matrices, 8 responses were calculated, see Fig. 1.S3J, K; responses for complex slow mode pairs were entirely similar) and used two metrics to quantify their resemblance to effective sum patterns. For each response, we calculate c_s, the correlation coefficient between the effective sum pattern and the response of the complex sum pair, and p_s, the proportion of the total response of the complex sum pattern pair in the direction of the effective sum pattern (equal to the dot product of the response of the complex sum pair with the effective sum pattern, each normalized to unit vector length). Fig. 1.S3L shows that c_s and p_s are indeed very high, demonstrating the equivalence of complex sum pattern pairs with single real sum patterns. Simulations of networks with complex pattern pairs show the same firing rates and correlation patterns as Fig. 1.2 (data not shown), further confirming the equivalence.

Section 6: Difference in correlation drop evoked by transient visual stimulation between the Bisley and Goldberg and the Falkner, Krishna et al. datasets

During the transient target visual response on target trials, there is a larger drop in correlation in BG than in FK, in both data (Fig. 1.1E-F) and model (Fig. 1.2E-F). During the transient distractor visual response on distractor trials, the correlation in the FK model rises to a higher level than the level to which the correlation drops in the corresponding period in the BG model (compare Fig. 1.2F to Fig. 1.2E), as is also seen in the data (compare Fig. 1.1F to Fig. 1.1E). In the model, these differences do not depend on whether the two local networks are coupled, but rather occur because the variation between the visual inputs to different neurons is smaller in the FK model than in the BG model, which was meant to roughly match the model firing rate variations to those observed in the data. BG had more visual response variations across cells than FK: distractor visual response standard deviations are 44 and 73 Hz for the two BG monkeys, and 29, 19, and 34 Hz for the three FK monkeys; target visual response standard deviations are 43 and 68 Hz for the BG monkeys, and 46, 26, and 48 Hz for the FK monkeys. The smaller visual input variation in the FK model compared to the BG model means that the weak Schur patterns are less activated relative to the dominant patterns, since the weak patterns are driven by variations in input across neurons while the dominant patterns are driven by mean inputs (Fig. 1.S6). Thus the dominant activity patterns are a larger component of the visual responses in FK, yielding the higher correlations. This finding suggests a prediction: in tasks or monkeys with smaller variations in visual response, this is due to smaller variations in visual input, which will manifest as higher correlations between target fixation activity and visual responses.

Section 7: Models of surround suppression without recurrent coupling in LIP

We consider any general change in external input to one local network that would suppress its firing. In addition to the scenario in the main text of withdrawing input to both E and I neurons, we can consider (1) addition of input to I cells alone; (2) withdrawal of input to E cells alone; (3) the combination of both (1) and (2). Scenario (1) mimics the input from the activated network to the suppressed network in the coupled network model (Fig. 1.2B), but lacks the reciprocal coupling by which the suppressed network in turn acts back upon the activated network.

We implemented these scenarios by modifying our inherited surround suppression model described in the main text. Whenever one local network receives visual or delay external input, the mean external input to the E cells and/or I cells of the other local network is changed according to the scenario being modeled, by an amount proportional to the mean visual or delay input. The change in input to each cell at time t is independently picked from a uniform distribution, whose mean is a fraction u of the mean visual and/or delay input at time t to the activated local network, and whose range is from 0 to twice its mean. For scenarios 1, 2, and 3, the respective u were chosen to be 1/20, 1/40, and 1/60 to produce magnitudes of suppression similar to that in the data.

Fig. 1.S5 illustrates these scenarios: while the changes in external inputs can be adjusted to produce changes in firing rates like those observed by FK (A-C), the correlation patterns do not resemble those observed by FK (D-F). In principle, if the changes in external input in any of the three scenarios is large enough, suppression of mean firing rate could be accompanied by a difference in the mean firing rates of the E

vs. I populations, which would result in a drop in correlation coefficient upon target onset on distractor trials, similar to the FK data. However, changes in external inputs need to be small to reproduce the subtle suppression of mean distractor trial firing rate induced by target onset in the FK data, and such small input changes do not change the relative activations of E and I populations enough to result in a drop in correlation (Fig. 1.S5).

Another possible scenario is that the interacting local networks of our FK model are actually in another area that we will call area Y; each local network of area Y projects to a corresponding local network in LIP in a manner such that the LIP local network inherits not only the mean firing rate over time in area Y, but also multi-neuronal activity patterns and therefore correlation patterns. This scenario is not impossible, but we consider it unlikely for three reasons. (1) Multi-neuronal firing patterns would be inherited if there is little convergence in the projections from area Y to individual LIP neurons (e.g., one-to-one connectivity). However, there is likely considerable convergence in intracortical projections and projections from subcortical areas to cortex, and thus the input an individual LIP neuron receives from a group of area Y neurons would reflect their average activity, regardless of the activity patterns across them. Highly variable weights at the area Y-to-LIP synapses could allow LIP to inherit area Y correlations to a certain extent, but not enough to reproduce the correlations observed by FK (data not shown). (2) The major areas projecting to LIP each show different response properties from LIP, suggesting that LIP activity patterns could not be simply inherited from them. Neurons in sensory areas such as MT and V4 fire weakly to small, stable visual stimuli such as the target present in the delay period of the FK task—they cannot account for reliable surround suppression in LIP by sustained saccade plans during the

delay. The projections from SC to LIP originate mainly from the superficial "visual" layers of SC which doesn't exhibit delay activity (Clower et al., 2001). In a recent experiment using a saccade task similar to the BG and FK tasks (Suzuki and Gottlieb, 2013), it was found that surround suppression in prefrontal cortex is much stronger than in LIP and exhibits qualitatively different properties. (3) The existence of local networks having one-dimensional dynamics, a prerequisite of our FK model, is well supported in LIP, but not in any other area. In conclusion, for the above reasons and for parsimony, we consider this scenario unlikely.

Section 8: Implications of different mechanisms of persistent activity on two-dimensional dynamics

In both our model and that of Ganguli et al. (2008), the LIP persistent activity during the delay period results from sustained top-down input from prefrontal cortex. This is a simplifying assumption, made because the focus of both studies was on the recurrent interactions within LIP. Possibilities for the actual mechanisms behind LIP persistent activity were discussed by Ganguli et al. (2008) in their Discussion. As they discussed in more detail, LIP is not likely to have attractor dynamics and sustain persistent activity by itself, since in the BG task, the strong visual response to the distractor is not able to trigger persistent activity. Therefore, they suggested ways whereby different oculomotor areas (LIP, FEF, dlPFC, SC, etc.) can recurrently interact with each other to generate distractor-resistant persistent activity in each area. One possibility is that each area acts as a "leaky attractor," and recurrently excites each other to balance out the leak so that each area has persistent activity. Or, one area might be able

to produce persistent activity by itself, but needs transient "gating" signals from other areas to be able to ignore distractors.

Because we do not have detailed knowledge of the connectivity between LIP and PFC, nor knowledge of the activity patterns across PFC neurons on the tasks we studied, attempts to include the recurrent interactions between LIP and PFC in our BG or FK models would be very under-constrained. However, we note that if persistent delay activity in LIP is generated through recurrent interaction with PFC, the conclusions of our study do not change. The dynamics of an LIP local network is still dominated by a small number of dominant patterns, but interaction with PFC effectively modulates the strength of self-excitation of the LIP dominant patterns, allowing them to be persistently active or decay based on the requirements of the task.

Section 9: The consequences of low-dimensional dynamics for attentional switching

In this section, we first examine the factors that determine the crossing time of the decaying distractor visual response and delay activity in FK, then examine the crossing of single neurons in both model and data.

The common crossing time of single neurons in BG can be explained by the one-dimensionality of LIP local dynamics around the time of the crossing (Ganguli et al., 2008). In state space, the multi-neuronal delay activity is a point on the one-dimensional line which is the direction of the slow mode, and the multi-neuronal visual response moves on this line towards the delay activity point as it decays. At the time that the multi-neuronal visual response meets the delay activity, the visual response of each neuron is equal to its delay activity, and thus this is the common crossing time.

In FK, the dynamics of an LIP local network are dominated by two activity patterns, the sum and difference patterns. If the two interacting local networks and their inputs are perfectly symmetric, then the activation of the sum pattern would be exactly the same on target and distractor trials, while that of the difference pattern would be exactly opposite. The activation of these patterns are evolving over time, but at each moment in the trial, sum activations are equal for the two trial types while difference activations are opposite. In this ideal case, after distractor offset, the decaying visual response on distractor trials and the delay activity on target trials only differ in their difference pattern activity.

We can approximate the dynamics of the difference pattern activity as follows:

$$\tau \frac{d}{dt} r_{diff} = -r_{diff} + \lambda_{diff} r_{diff} + I^{diff} \quad (3)$$

where is τ is the neuronal time constant, r_{diff} is the activity in the difference pattern, λ_{diff} is the eigenvalue of the difference pattern, and I^{diff} is external input to the difference pattern. For a given random instantiation of a global network, λ_{diff} is close to λ_D, the difference pattern eigenvalue of the mean population connectivity matrix, calculated in equations (1) and (2) from section 3 above. The difference pattern activity on distractor trials during the decay of the visual response is given by a solution to equation (3):

$$r_{diff}(t) = [(1 - e^{-t_0/\tau_{diff}}) \frac{I^{diff}_{visual}}{1-\lambda_{diff}}] e^{-t/\tau_{diff}} + (1 - e^{-t/\tau_{diff}}) \frac{-I^{diff}_{delay}}{1-\lambda_{diff}} \quad (4)$$

Here $r_{diff}(t)$ is the difference pattern activity as a function of time since the peak of the visual response, t_0 is the amount of time that visual stimulation was on, τ_{diff} is the time constant of the difference pattern and is equal to $\frac{\tau}{1-\lambda_{diff}}$, and I^{diff}_{visual} and I^{diff}_{delay} are

respectively the visual and delay input to the difference pattern (for clarity we define the inputs to be positive, and thus the negative sign before the delay input signifies that it drives the difference pattern negatively during distractor trials). The first term is the decaying visual response (the term within the bracket is the peak visual response reached during the transient visual stimulation, which is multiplied by an exponential decay term), while the second term is suppression due to delay input to the other local network. The difference pattern component of delay activity on target trials after distractor offset is simply the opposite of equation (4).

The crossing time T_c, the time when the decaying visual response and the delay activity are the same, is the time when they both have zero difference pattern activity, which is obtained by setting equation (4) to zero and solving for t:

$$T_c = \tau_{diff}\, ln\left(\frac{I_{delay}^{diff} + (1 - e^{-t_0/\tau_{diff}})\, I_{visual}^{diff}}{I_{delay}^{diff}}\right) \tag{5}$$

This shows that first, the crossing time is simply proportional to the time constant of the difference pattern. Second, logarithm term scales τ_{diff} and reflects the relative magnitude of the peak visual response and delay activity: the larger the peak visual response relative to delay activity, the longer it takes for the crossing to occur, and vice versa.

The above approximation depends crucially on the two local networks being symmetric, restricting two-dimensional dynamics to the single dimension of the difference pattern. However, as discussed in the Results section "Detailed analysis: two-dimensional dynamics explain correlation patterns," the stochastic components of connectivity and inputs means that the two local networks and the two trial types are not symmetric. This in turn means that sum pattern activities are not the same on the two trial

types, and difference pattern activities are not opposite. Thus the decaying visual response and delay activity evolve in a two-dimensional space instead of one dimension, and they do not meet in general. The crossing of their population PSTHs is not the crossing of their multi-neuronal activity patterns and not the common crossing of single neurons. The above analysis shows that although single neurons in FK do not act independently and are constrained by their network's two dominant activity patterns, this constraint is not as tight as in BG, and as a result we predict that single neurons crossing times would be more variable in FK than in BG.

Now we proceed to analyze the crossing dynamics of single neurons. We follow Bisley and Goldberg (2006) and Ganguli et al. (2008) and approximate the decaying distractor visual response with an exponential decay function (Fig. 1.S7A):

$$r_{visual}\, e^{-k\, t_c} \approx r_{delay} \qquad (6)$$

For each neuron, r_{visual} is its peak visual response to the distractor, k is its decay time constant, t_c is its crossing time, and r_{delay} is its delay activity at the time of crossing. Rearranging this gives:

$$ln\,(\, r_{visual} / r_{delay}\,) \;\approx\; t_c\, k \qquad (7)$$

Bisley and Goldberg (2006) and Ganguli et al. (2008) found that although $ln\,(\, r_{visual} / r_{delay}\,)$ and k each varies widely across neurons, these two quantities are highly correlated across neurons (Fig. 1.S7D). This is to say that t_c is approximately the same across neurons—the common crossing time.

Our prediction was that the linear relationship between $ln\,(\, r_{visual} / r_{delay}\,)$ and k found in BG would be less tight in the FK case, as shown by our model (Fig. 1.S7B-C). Indeed, that is the case in the FK data (Fig. 1.S7E). We also note that the FK response are

well-fit by exponentials (average R^2 across neurons and reward conditions: 0.97), similar to the BG data (Bisley and Goldberg, 2006).

Section 10: Contribution of random connectivity, inputs, and time constants to correlation patterns

In our main simulations (Fig. 1.2), we simulated multiple global networks and picked a random cell from each network to form a population. The networks each have random instantiations of connectivity, inputs, and time constants. To examine the contribution of each of these stochastic factors on correlation patterns, we did three sets of simulations where we modified our main FK simulation in the following way. For each of the three sets of simulations, we fixed one of the three factors across simulations, while still randomly instantiating the other two factors across simulation, and always picked the same cell to form our population. When we fixed connectivity and varied inputs and time constants across simulations, the correlation patterns broke down—instantaneous activity patterns throughout both types of trials were uncorrelated with fixation activity (data not shown). When inputs or time constants are the factors that were fixed across simulations, the FK correlation patterns remain (data not shown). These results arise because the FK correlation patterns are signatures of two-dimensional network dynamics, the key determinant of which is connectivity. Connectivity, by amplifying specific activity patterns (e.g. sum and difference patterns), gives rise to the correlations. Theses correlations, then, are necessarily correlations of patterns across different neurons, and would be absent when patterns of the same neuron across inputs are examined. On the other hand, the correlation patterns determined by network

dynamics (that is, by activity patterns strongly amplified by connectivity) are robust when the contributions from inputs and time constants are controlled.

2 Hebbian plasticity leads to biased representations in the lateral intraparietal area

2.1 Introduction

A recent study of the monkey parietal cortex uncovered a very surprising aspect of its neural responses. Fitzgerald et al. (2013) analyzed two associative learning experiments in which monkeys learned to group a set of visual stimuli into arbitrary categories. They found that after learning, neurons in the lateral intraparietal area (LIP) are category-selective, responding differently to stimuli belonging to different categories. Surprisingly, the representations of learned categories are overwhelmingly biased: while different categories are behaviorally equivalent, nearly all LIP neurons in a given animal prefer the same category.

LIP, the area in which the bias was observed, is an interface between brain regions subserving visual perception, eye movement planning, and visuospatial cognition, being reciprocally connected with visual sensory areas such as the middle temporal area (MT) and V4 as well as areas in the prefrontal cortex (PFC) such as the frontal eye field (FEF) and area 46 (REFs; check if the connections are all reciprocal). Neurons in LIP have visuospatial receptive fields (RFs), and it is well-known that they can encode visual attention and saccadic eye movement as well as variables concerning certain forms of decision making that have a visuospatial aspect to them (Bisley and Goldberg, 2010; Andersen and Cui, 2009; Gottlieb et al., 2009; Gold and Shadlen, 2007; Kable and Glimcher, 2009; Gottlieb, 2007).

More recently, LIP has been found to encode non-spatial, abstract factors. Freedman and Assad (2006), and Fitzgerald et al. (2011) trained monkeys on match-to-

category tasks (Fig. 2.1A), in which a number of different visual stimuli are grouped into arbitrary categories. After learning the task, LIP neurons were found to encode the category to which a stimulus belonged, in both their visual response to the stimulus, and in their sustained activity during a delay period after the presentation of the stimulus (Fig. 2.1B-E). In sensory areas, among neurons encoding a given stimulus variable, the neurons' preferred values of the encoded variable tend to evenly span the stimulus space (V1: Hubel and Wiesel, 1962; MT: DeAngelis and Uka, 2003; V4: Conway and Tsao, 2009; V4: Hegde and Van Essen, 2007; V4: Lehky et al., 2011). Therefore, the finding that almost all recorded LIP neurons prefer the same category was very unexpected (Fitzgerald et al., 2013). A similarly striking bias was also observed in Bennur and Gold (2011).

Here we propose a model in which LIP develops category-selectivity and representational bias under the guidance of PFC. Given plastic synapses from PFC to LIP and from MT and V4 to LIP, we show that recurrent connectivity in LIP allows simple Hebbian plasticity to give rise to circuits that explain a variety of experimentally observed phenomena. In the following text, we examine the experimentally observed category-selectivity and bias, lay out considerations that lead to the premises of our model, reproduce experimental findings through simulations, present mathematical analyses of the model that give intuitions for its underlying mechanisms, and offer predictions for experiments, one of which we also test.

2.2 Results

Category-selectivity and representational bias in LIP

During a trial of the match-to-category task (Fig. 2.1A), the LIP population exhibits a visual response during the presentation of the sample stimulus and a sustained response during the delay period after sample stimulus offset (Fig. 2.1B-E). During both the sample visual response and the delay response, LIP fired at different rates during trials in which sample stimuli belonging to different categories were shown. Importantly, single LIP neurons robustly encoded the category to which stimuli belonged, but not the identity of stimuli (Freedman and Assad, 2006).

The fact that category-selectivity can be observed in the averaged response of thee population (Fig. 2.1E) suggest that the neural representations of different categories are unequal. In fact, examining on a cell-by-cell basis, we see that for a given monkey under a given task condition, most recorded LIP neurons tend to have the same order of preference for the different categories (Fig. 2.1F). In general, the bias is stronger during the delay period than during the sample period, i.e. more neurons share the majority preference during the delay than the sample period. Furthermore, Fitzgerald et al. (2013) also observed a bias in the data of Bennur and Gold (2011). Bennur and Gold trained monkeys to report the direction of noisy arrays of moving dots, where leftward or rightward motion was reported by saccades to a red or green target, respectively. In two monkeys, LIP neurons overwhelmingly preferred rightward motion and red target over leftward motion and green target, during both the presentation of the motion stimuli and the delay period after their offset.

Premises for a model of bias development in LIP

The biases of category representation in LIP are results of learning. The monkeys

only experienced the stimuli as members of categories when they started learning the tasks; the categories are nonexistent before. Fig. 2.2A shows the rank ordering of LIP single neuron responses when passively viewing the same motion stimuli as used in the match-to-category task, recorded from monkey H before it was trained on that task. Indeed, LIP neurons did not exhibit selectivity or representational bias for the "categories" which were not yet defined at that time.

The naive external connectivity and internal circuitry of LIP underlie its classical visuospatial response, but how does it respond to abstract, nonspatial factors? Balan and Gottlieb (2009) inactivated LIP during three tasks in which LIP encoded three nonspatial factors: limb motor planning, time estimation, and reward expectation. LIP inactivation produced visuospatial deficits, but had no effect on the three nonspatial aspects of performance. This suggests that LIP does not play a major functional role in processing these nonspatial factors; instead, LIP may be simply reflecting the nonspatial signals in its external inputs.

Where do the external inputs carrying nonspatial signals come from? One likely candidate is PFC, an area with reciprocal connections with LIP and known to encode abstract, cognitive factors. Recordng during the same match-to-category task as Freedman and Assad (2006), Swaminathan and Freedman (2012) found that PFC exhibited similar category-selectivity as LIP. Moreover, there is little or no bias in the category representation in PFC (S. Swaminathan, personal communication).

Given these information, we make the following assumptions, laying out the premises of our model. As a monkey is being trained on the match-to-category task and learns to group stimuli into categories, PFC first acquires category selectivity, with

approximately equal number of neurons preferring each category. As training progresses, through plasticity at the PFC-to-LIP and MT-to-LIP synapses, under conditions to be discussed below, LIP acquires both category selectivity and representational bias (Fig. 2.2D-E).

Architecture and mechanisms of the model

We first present the most basic version of our model and show how it reproduces bias during the delay period of the motion category task, illustrating the general mechanisms underlying this and fuller versions of the model. We model a network of interconnected LIP neurons. In addition to recurrent connections from within LIP, each neuron also receives bottom-up projections from area MT and top-down projections from PFC (Fig. 2.2C). Model MT neurons respond to visual motion stimuli and are direction-selective. We model two PFC populations with equal number of neurons, each preferring one of the two motion categories. To illustrate the model mechanisms, we first model all-to-all connectivity from PFC to LIP.

We model LIP activity with a standard linear rate equation:

$$\frac{dr}{dt} = -r + Wr + Sh \tag{1}$$

Here, *r* is a vector of the response of LIP neurons, *W* is the LIP recurrent synaptic weight matrix, *S* is the external input synaptic weight matrix, and *h* is the vector of external inputs. During the delay period, no stimulus is shown and MT is silent, so we can model PFC inputs as the sole external inputs.

LIP delay activity is then the steady-state activity to PFC persistent input:

$$r = (I - W)^{-1} Sh = \tilde{W} Sh \tag{2}$$

where we have defined $\tilde{W} = (I - W)^{-1}$.

In our model the only plastic synapses are the ones onto LIP neurons from external inputs, that is, *S*. We initiate *S* with weak, random weights, and implement hard lower and upper bounds on the weight of each synapse. We use a standard Hebbian plasticity rule to model the development of *S*:

$$\tau \frac{dS}{dt} = < rh^T > - < r >< h^T > \tag{3}$$

where τ is the plasticity rate, and <•> denotes averages over all trials regardless of the stimulus presented. Thus, synapses potentiate (or depress) when presynaptic firing in PFC and postsynaptic firing in LIP are correlated (or anti-correlated). Plugging (2) into (3), we obtain:

$$\begin{aligned} \tau \frac{dS}{dt} &= \tilde{W} S(< hh^T > - < h >< h^T >) \\ &= \tilde{W} SH \end{aligned} \tag{4}$$

where we have defined the input correlation matrix $H =< hh^T > - < h >< h^T >$. (4) is a linear differential equation, which can be understood through eigendecomposition of the matrices $\tilde{W}$ and *H*. Specifically, the change in *S* can be decomposed into change in the directions of $V_{\tilde{W},i} V_{H,j}^T$ (a matrix the same size as *S*), the outer product of the *i*th eigenvector of $\tilde{W}$, $V_{\tilde{W},i}$, and the *j*th eigenvector of *H*, $V_{H,j}$. The largest change in *S* occurs in the direction where $\alpha_{\tilde{W},i} \alpha_{H,j}$ has the most positive real part, where $\alpha_{\tilde{W},i}$ and $\alpha_{H,j}$ are the eigenvalues associated with $V_{\tilde{W},i}$ and $V_{H,j}$,

respectively.

Thus, to understand the structure of change in *S*, we first examine *H* (Fig. 2.3B-D), followed by $\tilde{W}$. Using h^i to denote the inputs on trials where category *i* is presented,

$$\begin{aligned} H &= \frac{1}{2}(< h^1 h^{1T} > + < h^2 h^{2T} >) - \frac{1}{2}(< h^1 > + < h^2 >)\frac{1}{2}(< h^{1T} > + < h^{2T} >) \\ &= \frac{1}{2}(C^1 + C^2) + \frac{1}{4}(< h^1 >< h^{1T} > + < h^2 >< h^{2T} >) - \frac{1}{4}(< h^1 >< h^{2T} > + < h^2 >< h^{1T} >) \end{aligned} \quad (5)$$

Here C^i denote the input covariance matrix on trials where category *i* is presented. Assuming that the responses of different PFC cells to the same category are independent, we can simplify H:

$$H = \frac{1}{2}(\sigma_p^2 + \sigma_{np}^2)I + \frac{1}{4}(< h^1 >< h^{1T} > + < h^2 >< h^{2T} >) - \frac{1}{4}(< h^1 >< h^{2T} > + < h^2 >< h^{1T} >) \quad (6)$$

where σ_p^2 and σ_{np}^2 denote the variance across PFC cells of response to the preferred and non-preferred categories, respectively, and *I* denotes the identity matrix. Using m_p and m_{np} to denote the mean PFC response to preferred and non-preferred categories, respectively, and *A* to denote the square matrix of all ones with half the number of rows and columns as *H*, we can write

$$H = \frac{1}{2}(\sigma_p^2 + \sigma_{np}^2)I + \frac{1}{4}(m_p - m_{np})^2 \begin{pmatrix} A & -A \\ -A & A \end{pmatrix} \quad (7)$$

The second term expresses the anti-correlation of responses of the two PFC populations preferring different categories. This results in the dominant effect of *H* on the development of *S*: for a given LIP cell, all synapses from PFC cells preferring one category tend to increase in strength, and all synapses from PFC cells preferring the other category tend to decrease in strength. This gives rise to category selectivity for each LIP cell.

Now we turn to examine $\tilde{W}$. Since $\tilde{W} = (I - W)^{-1}$, their eigenvectors are the same, and the eigenvalues of $\tilde{W}$ and W have a one-to-one correspondence: $\alpha_{\tilde{W},i} = (1 - \alpha_{W,i})^{-1}$, where $\alpha_{\tilde{W},i}$ and $\alpha_{W,i}$ are the ith eigenvalues of $\tilde{W}$ and W, respectively. We assume that all eigenvalues of W have real parts less than 1, so that LIP dynamics [equation (1)] are stable. Then, the rank ordering of the eigenvalues of $\tilde{W}$ and W by their real parts are the same. Thus, we can understand the effects of $\tilde{W}$ on synaptic development by examining the leading eigenmodes of W (Fig. 2.3E-G). If the LIP recurrent connectivity has stronger excitation than inhibition, then the eigenvector with the most positive eigenvalue is one where most elements have the same sign (Fig. 2.3G). This means that LIP neurons tend to activate together, thus causing each other to develop the same category selectivity, giving rise to biased representation.

Intuitively, the effects of PFC input correlations and LIP recurrent connectivity on synaptic development can be summarized as follows. Consider a single LIP cell before training, which receives weak and roughly equal inputs from PFC cells encoding categories 1 and 2. During early training, the random synaptic connections from the two PFC populations and their noisy firing could lead to PFC population 1 activating the LIP cell slightly more than PFC population 2. This would lead to slight potentiation of the population 1 synapses onto the LIP cell, and a slight depression of the population 2 synapses onto it. Because of these synaptic changes, later on in training, PFC population 1 would activate the LIP cell more, leading to more potentiation of their synapses. Conversely, the synapses from PFC population 2 would depress. This cycle of positive feedback eventually would lead to large weights at population 1 synapses and small

weights at population 2 synapses onto this LIP cell, making it prefer category 1. Now consider other LIP cells in its neighborhood. When the one LIP cell first starts to weakly prefer category 1, it excites other LIP cells in its vicinity, biasing them to also fire more in response to category 1. This leads to potentiated synaptic weights from PFC population 1 to this group of LIP cells, which recurrently excite each other to further increase their response to category 1. This cycle of positive feedback across LIP cells eventually leads to most of them preferring the same category.

Development of bias after re-definition of categories

Freedman and Assad (2006) have trained monkeys on the match-to-category task, with 12 directions belong to 2 categories. After the monkeys learned the categories, biased representations developed in LIP for the categories (Fig. 2.4A and C). Then, the categories were re-defined, such that the new category boundaries are orthogonal to the old ones, dividing the 12 directions into 2 new categories (Fig. 2.4A-D). Since the old and new category boundaries are orthogonal, the old category representation in LIP would translate to neither category selectivity nor category bias. However, after the monkeys learned the new categories, neural representation in LIP changed, such that neurons have both category selectivity and biased representation for the new categories (Fig. 2.4B and D).

To model this phenomenon, we made further assumptions on the neural representation in PFC as animals learn new category definitions (Fig. 2.4E). We assume that before category re-training, PFC contains two equal neural populations preferring the two categories, and a third population that do not encode categories—neurons in all three

populations have similar projection patterns onto LIP neurons. After animals have learned the new categories, the neural representation in PFC has effectively been shuffled, such that a random subset of neurons encoding the old categories stops encoding categories, and another random subset of neurons not previously encoding the old categories starts to encode the new categories. As the old and new category boundaries are orthogonal, for a given cell that encodes categories after re-training, we randomly assign it a category preference, regardless of its category preference before re-training.

Training on the first set of categories resulted in category selectivity and bias in this model, as expected (Fig. 2.4F). For the re-training on new categories, the change in PFC representation changes the input correlation matrix *H*, which drives LIP neurons to again develop category selectivity and bias for the new categories (Fig. 2.4G).

Development of bias for three categories of shapes

In another set of match-to-category experiments, Fitzgerald et al. (2011) trained monkeys to categorize six abstract shapes into three categories (Fig. 2.5A-B). In both monkeys, category selectivity developed in LIP cells, while a category bias also developed in one of the monkeys (Fig. 2.5A).

To model LIP learning in this task, we assume that the sensory inputs encoding the shapes come to LIP from V4, where each LIP cell receives inputs from a set of V4 cells preferring a random set of shapes. Furthermore, we assume that after the animals learn the categories, PFC contains three equal populations of cells preferring each of the three categories, which send equal projections to LIP (Fig. 2.5C).

In this case, the input correlation structure is similar to that analyzed above (Fig. 2.3B-D), where inputs from PFC cells preferring the same category are correlated, while inputs from PFC cells preferring different categories are anti-correlated. The effect of this input correlation structure on PFC-to-LIP synaptic plasticity leads to LIP cells developing differential responses to the three categories, making them category-selective, as in the direction category tasks above. The recurrent connectivity in LIP again leads to LIP cells developing the same category selectivity, resulting in biased representation (Fig. 2.5D).

Spatial clustering of LIP category preferences

In our model, LIP cells give feedback to each other during learning via their excitatory recurrent connections, making mutually connected cells develop the same category preference. Since spatially nearby cells are more heavily recurrently connected, this suggests there might be spatial clustering of category preferences. To examine this, we modified our model of the LIP network, from being situated on a 1D ring to a 2D cortical surface. The LIP cortical surface contains rough topological maps of visual space: neurons whose RFs are nearby are more likely to be located close to each other on the cortical surface (Blatt et al., 1990; Patel et al., 2010). In our simple model, there is an exact topological correspondence between neurons' location on the 2D cortical surface and their RF positions. Thus, the connection probability between neurons is a decreasing function of the distance between their RF positions. In this model, category selectivity and bias develops as in the 1D model. On simulations where the bias is weak, we indeed observe spatial clustering of category preferences (Fig. 2.6A-B).

We wish to test our model prediction that cells with the same category

preferences would have spatially clustering RFs. The Fitzgerald et al. (2013) datasets are not informative in this regard because the bias in those datasets are too strong, where almost all cells have the same category preference. Thus we turn to an experiment by Oristaglio et al. (2006), which exhibits relatively weaker biases of response preferences in LIP. Briefly, their behavioral task is as follows. A monkey fixates a central fixation spot and holds two bars with its two hands. There are four figure-8s on the screen, and after ~500 ms of fixation, two line segments are removed from each figure-8. This results in one of the four stimuli turning into either a leftward-facing or rightward-facing letter "E", which is the task-relevant cue for the monkey, while the other stimuli turn into task-irrelevant distractors. The monkey is rewarded if it releases its right hand when the "E" is rightward-facing and releases its left hand when the "E" is leftward-facing, while maintaining fixation. In this task, LIP neurons can encode which hand is being used to release a bar. This hand selectivity is dissociated from selectivity for the identity of the stimulus or the position of the hand (Oristaglio et al., 2006).

This hand selectivity is similar to the category selectivity observed by Fitzgerald et al. (2013) in that they are both "non-spatial" response properties, as opposed to the classic spatial responses of LIP neurons, including visual response, delay response, and saccadic responses. Furthermore, hand selectivity is likely learned through training, like category selectivity. Thus, we hypothesize that hand selectivity develops in LIP through the same mechanisms as we proposed for category selectivity. Thus, we expect hand preference of LIP cells to show spatial clustering. Indeed, this is what we observed (Fig. 2.6C).

Biased representation of continuous stimulus variables

Fitzgerald et al. (2013) suggested that biased representation might be a neural strategy for encoding discrete stimulus variables, such as categories, and that LIP neurons might encode continuous stimulus variables with more unbiased stimulus preferences. However, the learning mechanisms we proposed is compatible with LIP learning biased representations of continuous stimulus variables. In tasks where animals compare the magnitude of continuous stimulus variables, neurons can encode the stimulus magnitude. For example, Ferrera et al. (2009) trained animals to compare the speed of moving dots, and found two roughly equal populations of FEF neurons with monotonic tuning functions for speed, with one population preferring faster speeds and another preferring slower speeds. If these prefrontal neural populations such as these guide LIP learning as we proposed, we predict that LIP would develop biased representations for the continuous variable encoded by the prefrontal neurons (Fig. 2.7). For example, almost all LIP neurons could prefer large magnitudes of the continuous stimulus variables.

2.3 Discussion

Our model suggests that biased representation develops in LIP due to its net excitatory recurrent connectivity. We believe the ability of LIP to develop biased representations is indicative of its wider role in visuospatial cognition. Specifically, this role of LIP is to learn to encode the significance of visual stimuli, in order to guide attentional allocation and eye movements. We elaborate on this idea below.

The biased representation of abstract categories examined here are not intuitive—it is not clear why LIP neurons would all prefer one non-spatial stimulus variable over

others. However, similar "biases" that LIP neurons exhibit are intuitive and taken for granted. For example, in tasks where different visual stimuli are associated with different amounts or probabilities of reward, almost all LIP neurons prefer stimuli predicting larger amounts of reward (Coe et al., 2002; Dorris and Glimcher, 2004; Sugrue et al., 2004; Peck et al., 2009). Similarly, almost all LIP neurons prefer task-relevant stimuli over irrelevant ones (Oristaglio et al., 2006). We suggest that such "biases" for reward-predicting stimuli could have developed in LIP through the same mechanisms that we proposed for the development of biased category representations. These mechanisms allow LIP to identify behaviorally significant visual stimuli, allowing attention to be directed towards them covertly or overtly.

Interestingly, in FEF and dlPFC, areas that we hypothesize guide LIP in learning the significance of stimuli, a small majority of neurons prefers stimuli associated with higher reward (Leon and Shadlen, 1999; Coe et al., 2002; Roesch and Olsen, 2003; Pan et al., 2008; Kennerly and Wallis, 2009; Teichert et al., 2014). That is, PFC contains neural populations preferring stimuli predicting larger rewards as well as preferring smaller rewards. It's possible that this is because PFC is involved with encoding all stimuli regardless of their current behavioral significance, allowing flexible behavior in environments where the significance of stimuli can change. Through plasticity at the PFC-to-LIP synapses, LIP is able to learn the current significance of stimuli from PFC and direct attentional processes accordingly.

2.4 Methods

LIP connectivity

We modeled a LIP network on a 1D ring, with N cells, half excitatory (E) and half inhibitory (I). Note in all our model networks of LIP, PFC, MT, and V4, positions on the ring correspond to the visual location of the RFs, not the preferred motion direction of the cell. A pair of E and I cells resides at each of $N/2$ positions on the ring. The connection probability from one cell to another is $\exp(-d/c)$, where d is the distance between the two cells on the ring, and c is a parameter describing the range of connections; we do not allow autapses. The weights of each E-to-E, E-to-I, I-to-E, and I-to-I synapse are w_{EE}, w_{IE}, w_{EI}, and w_{II}, respectively. After each probabilistically connected weight matrix is generated, we scale the matrix by dividing each element by 1.1 times the most positive real part of the eigenvalues of the matrix, in order to prevent instability.

Our model of a LIP network on a 2D surface is similar, except that there is a square grid of N_s by N_s positions with a pair of E and I cells at each position.

Top-down and bottom-up connectivity

Top-down and bottom-up connections together form the input connectivity matrix S. We model top-down input to LIP from N_f E cells in PFC, also existing on a 1D ring. At each position in the PFC ring, there are equal numbers of cells preferring each category. There is topological connectivity from PFC to LIP, with each PFC cell projecting to n_{td} LIP cells with positions on the LIP ring nearest to the corresponding position of the PFC cell on its ring. For the model of the category re-definition experiment, there are equal number of category-encoding and non-encoding cells in PFC, and a fraction of f_{re} of the cells encoding the old categories do not encode the new categories.

We model bottom-up input to LIP from MT for the direction category

experiments or from V4 for the shape category experiment. The bottom-up inputs are E cells also on a 1D ring, with equal numbers of cells preferring each stimulus at each position on the ring. As with the top-down inputs, each bottom-up input cell project topologically to n_{bu} LIP cells. For each LIP cell, we randomly select from 0 up to n_{trim} stimuli that do not give input to that cell.

Synaptic plasticity and LIP responses

We model the change in the input connectivity matrix S with a standard Hebbian plasticity rule [equation (4) in Results]. Each nonzero weight in S is initially independently drawn from a uniform distribution on (0, 0.1). During learning, each weight is bounded by 0 and 1. Before and after learning, LIP responses are calculated with equation (2) in Results.

Parameters

$N = 606$, $c = 15$, $w_{EE} = 1.3$, $w_{IE} = 1.1$, $w_{EI} = -1$, $w_{II} = -1.1$, $N_s = 30$, $N_f = 152$, $n_{td} = 26$, $f_{re} = 0.25$, $n_{bu} = 10$, $n_{trim} = 4$.

Processing of the Oristaglio et al. (2006) data

For each cell, spike counts from -200 to 0 ms before bar release on left hand release trials were compared with those on right hand release trials, using a two-sample t test. A cell with a p value of less than 0.05 was considered hand-encoding.

Figure 2.1

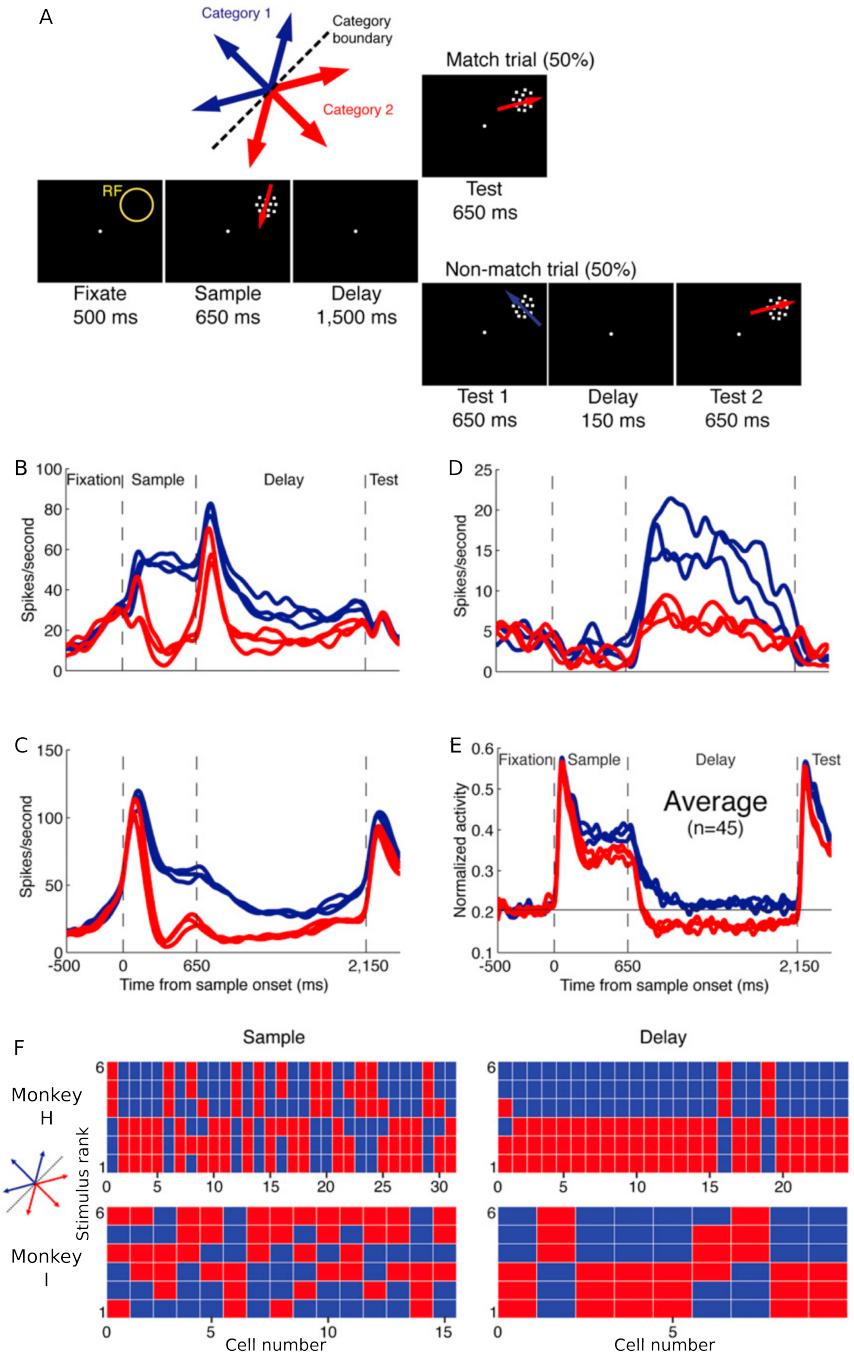

Figure 2.1
LIP shows biased representation of abstract categories.
(A) The match-to-category task. After fixation, an array of coherently moving dots is shown as the sample stimulus inside the receptive field (RF). On a given trial, the sample stimulus could be moving in one of six directions, which are grouped into two categories. The sample stimulus is followed by a delay, after which another array of coherently moving dots is shown as the test stimulus. If the sample and the test stimuli belong to the same category, the monkey is required to release a bar that it has been manually holding

to receive a juice reward. If they belong to different categories, the monkey is required to continue holding the bar until another test stimulus appears that belongs to the same category as the sample.
(B-D) PSTHs of three example cells from monkey H. In each panel, each of the six traces denote trials in which one of the six sample stimuli was shown, colored according to the category to which the stimulus belongs.
(E) Population-averaged, max-normalized PSTH from monkey H.
(F) Rank ordering of neural responses during the sustained sample (200-650 ms after motion onset; left panels) and late delay (750-1,500 ms after motion offset; right panels) periods for monkey H (top panels) and monkey I (bottom panels). In each panel, each column denotes a cell, and the six small rectangles denote the six sample stimuli (colored according to their categories), arranged such that the stimulus that evoked the highest response for that cell is at the top of the column, and the stimulus that evoked the second highest response is second from the top, etc. Thus, during the delay period for monkey H (top-right panel), for example, almost all cells prefer the blue category over the behaviorally equivalent red category.

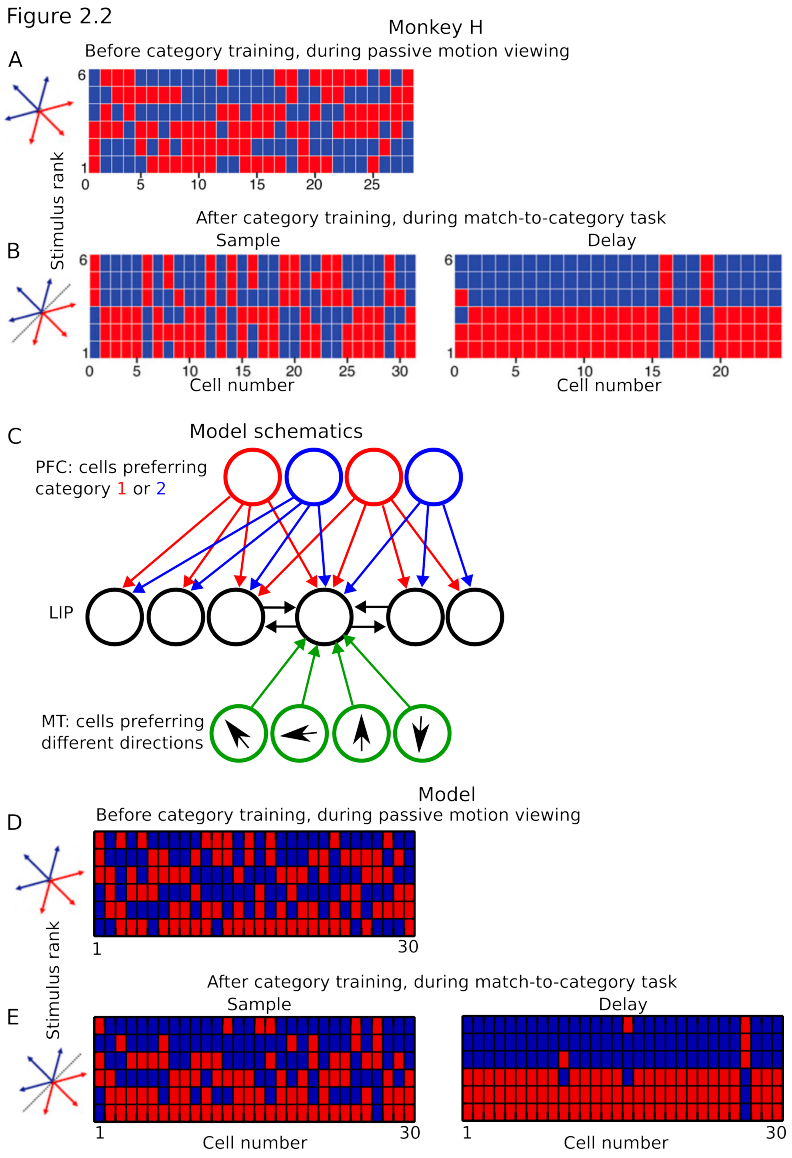

Figure 2.2
Model of category learning and development.
(A) The selectivity of LIP cells from monkey H to the sample stimuli during passive viewing, before they were grouped into categories through training on the match-to-category task (Fig. 2.1A). Same format as Fig. 2.1F.
(B) The selectivity of LIP cells from monkey H after training. Replotted from the top panels of Fig. 2.1F.
(C) Model schematics. We assume that during training on the match-to-category task, PFC first learns the categories, and through plasticity at the PFC-to-LIP and the MT-to-LIP synapses, LIP cells acquire both category selectivity and biased representation. Equal numbers of PFC cells prefer categories 1 and 2, and they project equally to LIP cells. LIP cells are recurrently connected with each other, and each in addition receives input from MT neurons with a random set of preferred directions. See text for details.
(D-E) Same as A-B, but from the model.

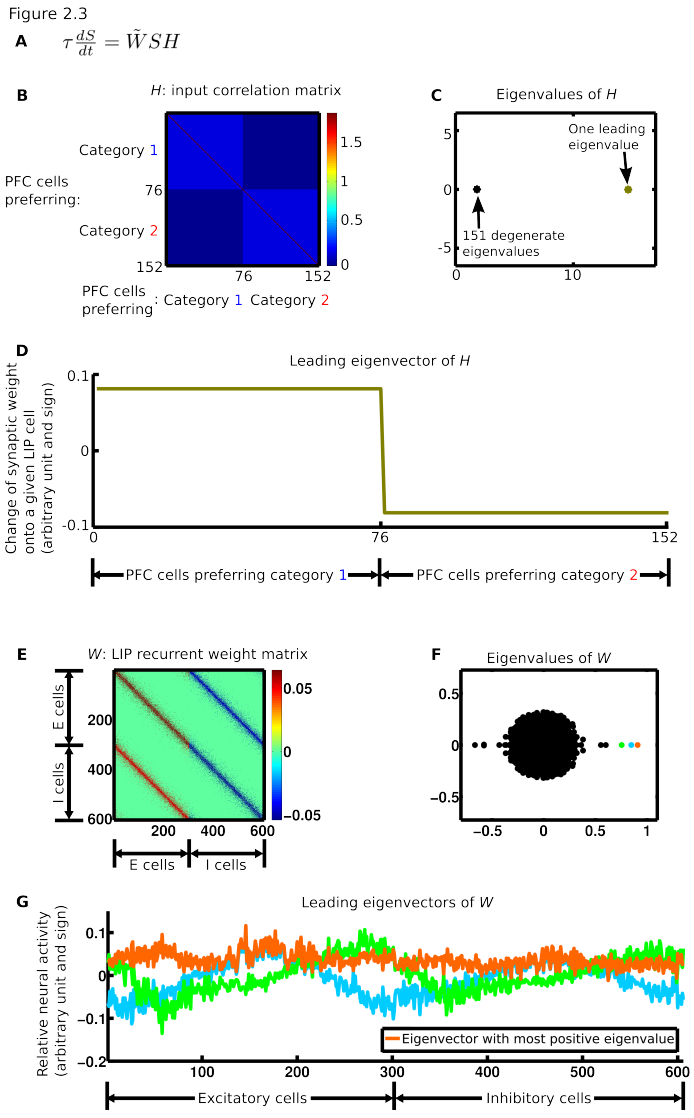

Figure 2.3

The mechanisms of category learning and bias development in a reduced model.
(A) The development of the synaptic weights S depends on H, the correlation of the inputs to LIP from PFC cells, and $\tilde{W}$, a matrix describing the effects of the recurrent connectivity in LIP. See text for derivation.
(B) The input correlation matrix H.
(C) The eigenvalue spectrum of H shows one dominant eigenvalue.
(D) The leading eigenvector of H: the elements corresponding to PFC cells preferring the same category have the same sign, while elements corresponding to PFC cells preferring different categories have opposite signs. Thus, for a given LIP cell, PFC input correlations tend to cause all synapses from PFC cells preferring one category to increase in strength, and all synapses from PFC cells preferring the other category to decrease in strength. This gives rise to category selectivity for each LIP cell.
(E) The LIP recurrent connectivity matrix W. Here LIP is modeled as a 1D ring, where the recurrent connection probability decreases with distance on the ring. Since

$\tilde{W} = (I - W)^{-1}$, the eigenvalue spectra of $\tilde{W}$ and W have the same structure, while their eigenvectors are the same.
(F) The eigenvalue spectrum of W.
(G) The leading eigenvectors of W. For the eigenvector with the most positive eigenvalue, most elements have the same sign. This means that LIP neurons tend to activate together, thus causing each other to develop the same category selectivity, giving rise to biased representation.

Figure 2.4

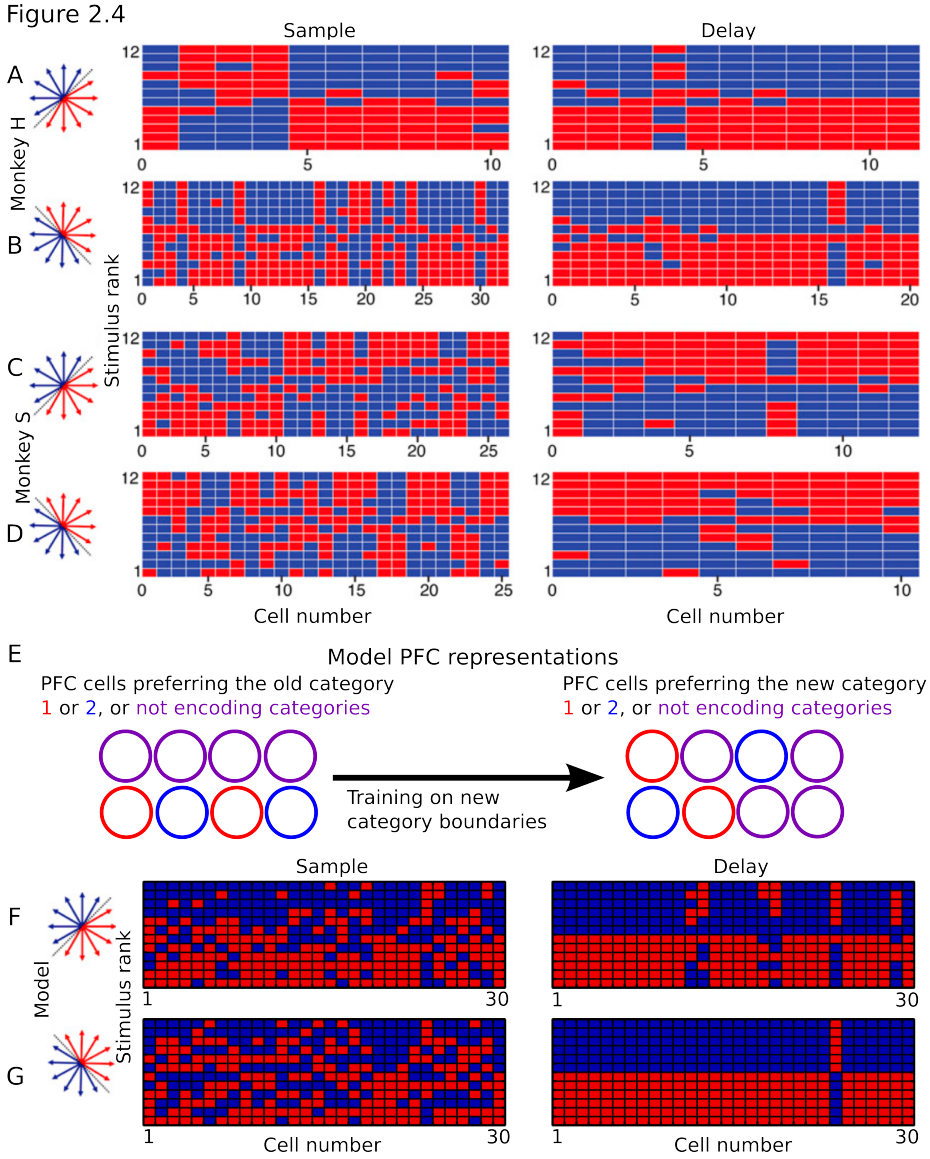

Figure 2.4
Model reproduces emergence of biased representation after categories were redefined. (A-D) Monkeys H and S were first trained to group 12 motion directions into two categories, and LIP developed biased representations of the categories (A and C). The monkeys were then trained to group the same motion stimuli into two new categories, and LIP again developed biased representations of the redefined categories (B and D). (E) Our model assumes that after training on the first category definition, PFC has an equal number of neurons encoding each category, and another population that does not encode categories. All PFC neurons project to LIP. Learning a new category definition changes the coding in PFC, and a random set of neurons encode the new categories. (F-G) Our model reproduces biased representations of a first set of categories (F), as well as re-development of bias after re-definition of the categories (G).

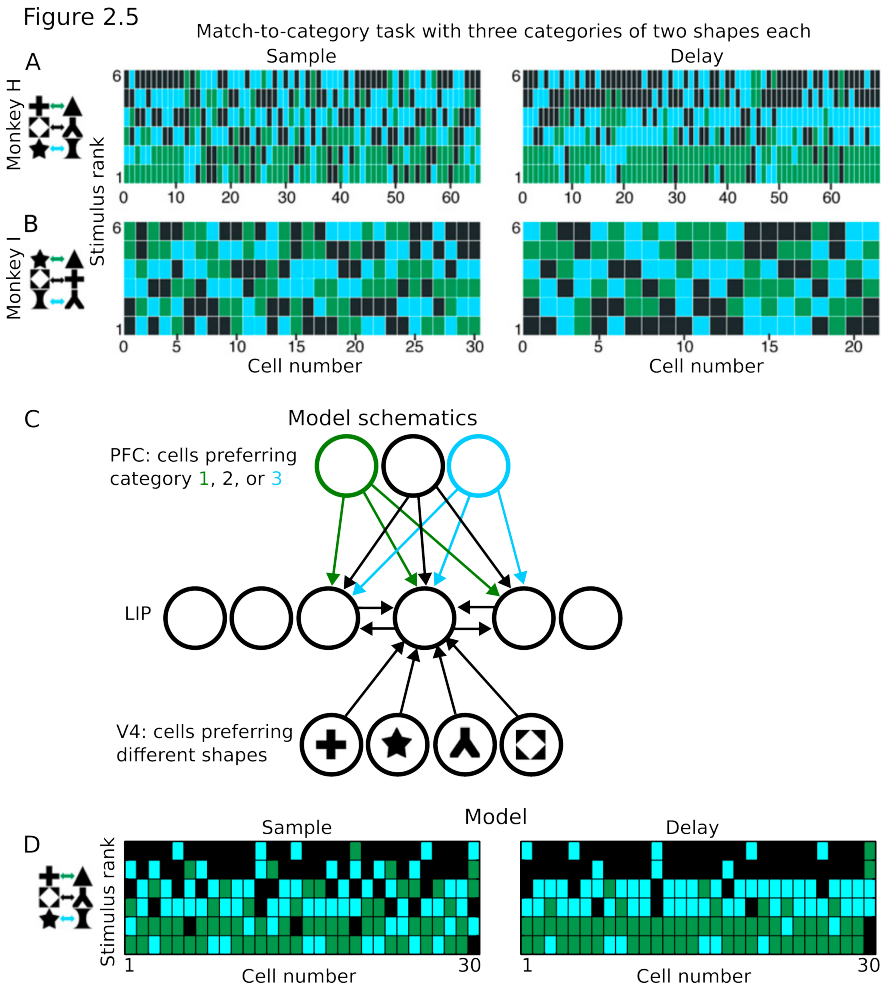

Figure 2.5

Model reproduces the biased representation of three categories of abstract shapes.
(A-B) Monkeys H (A) and I (B) were trained on a match-to-category task with identical structure as Fig. 2.1A but different stimuli and categories: six abstract shapes grouped into three categories. Monkey H (A) develops a biased representation in this task.
(C) Model schematics. Same as the model in Fig. 2.2C, except that there are three equal PFC populations encoding the three categories, and that the bottom-up inputs to LIP in this task come from V4 shape-selective neurons.
(D) Model reproduces biased representation. Note that on some simulations (i.e. for some random instantiations of connectivity and inputs), the model does not converge to a biased representation (data not shown), as is the case for Monkey I (B).

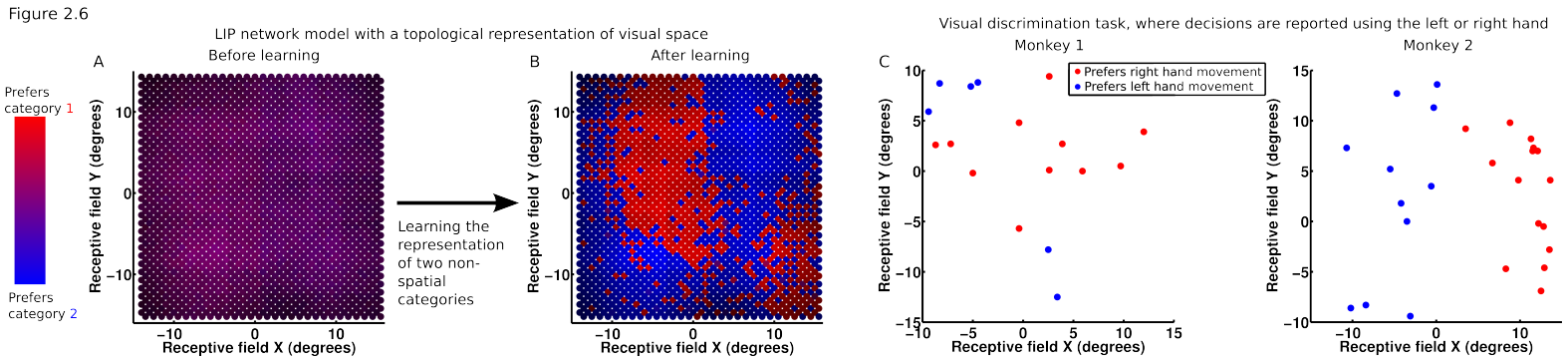

Figure 2.6
Data confirms model prediction that a patch of connected neurons tends to develop the same category selectivity.
(A-B) We extended our model of LIP from a 1D ring to a 2D surface, where neurons have a topological representation of visual space. Connection probability between LIP neurons falls off with distance on the surface, and thus with distance between their RFs in visual space. Category selectivity before (A) and after (B) category training shows that patches of cells with the same category preference develop.
(C) In a visual discrimination task where decisions are reported using either the left or right hand, Oristaglio et al. (2006) found that LIP cells have hand selectivity independent of visuospatial information. Plotting the hand preference of LIP cells as a function of their RF positions confirms our model prediction that cells with the same preference would have RFs that cluster in space.

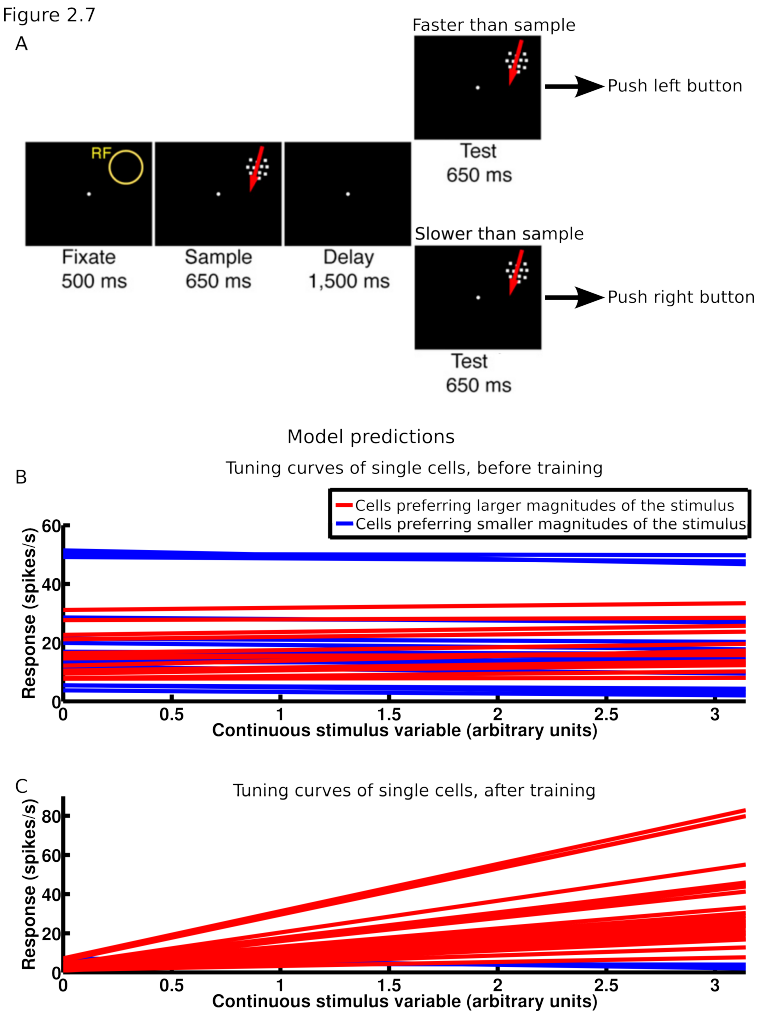

Figure 2.7
Model predicts biased representation of continuous stimulus variables.
(A) A speed comparison task modified from the match-to-category task in Fig. 2.1A. All sample and test stimuli are coherent dot motion in the same direction; the monkey is required to compare the speed of the sample and test stimuli.
(B-C) Our model predicts that LIP would develop biased representations of continuous stimulus variables, such as motion speed in the task in A, where almost all cells would prefer larger magnitudes or almost all cells would prefer smaller magnitudes of the variable.

References

Allman, J., Miezin, F., and McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. Annu. Rev. Neurosci. *8*, 407–430.

Andersen, R. A, and Cui, H. (2009). Intention, action planning, and decision making in parietal-frontal circuits. Neuron *63*, 568–583.

Baizer, J.S., Ungerleider, L.G., and Desimone, R. (1991). Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. J. Neurosci. *11*, 168–190.

Balan, P.F., Oristaglio, J., Schneider, D.M., and Gottlieb, J. (2008). Neuronal correlates of the set-size effect in monkey lateral intraparietal area. PLoS Biol. *6*, e158.

Bennur, S., and Gold, J.I. (2011). Distinct representations of a perceptual decision and the associated oculomotor plan in the monkey lateral intraparietal area. J Neurosci *31*, 913–921.

Bisley, J.W., and Goldberg, M.E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. Science *299*, 81–86.

Bisley, J.W., and Goldberg, M.E. (2006). Neural correlates of attention and distractibility in the lateral intraparietal area. J. Neurophysiol. *95*, 1696–1717.

Bisley, J.W., and Goldberg, M.E. (2010). Attention, intention, and priority in the parietal lobe. Annu. Rev. Neurosci. *33*, 1–21.

Blatt, G.J., Andersen, R.A., and Stoner, G.R. (1990). Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area LIP) in the macaque. J. Comp. Neurol. *299*, 421—445.

Born, R.T., and Bradley, D.C. (2005). Structure and function of visual area MT. Annu. Rev. Neurosci. *28*, 157–189.

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. Neuron *68*, 362–385.

Carandini, M., and Heeger, D.J. (2012). Normalization as a canonical neural computation. Nat. Rev. Neurosci. *13*, 51–62.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Clower, D.M., West, R.A., Lynch, J.C., and Strick, P.L. (2001). The inferior parietal lobule is the target of output from the superior colliculus, hippocampus, and cerebellum. J. Neurosci. *21*, 6283–6291.

Coe, B., Tomihara, K., Matsuzawa, M., and Hikosaka, O. (2002). Visual and anticipatory bias in three cortical eye fields of the monkey during an adaptive decision-making task. J. Neurosci. *22*, 5081–5090.

Constantinidis, C., and Wang, X.-J. (2007). A Neural Circuit Basis for Spatial Working Memory. Neuroscientist *10*, 553–565.

Conway, B.R., and Tsao, D.Y. (2009). Color-tuned neurons are spatially clustered according to color preference within alert macaque posterior inferior temporal cortex. Proc. Natl. Acad. Sci. U. S. A. *106*, 18034–18039.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500-1509.

Dayan, P., and Abbott, L.F. (2005). Theoretical Neuroscience (Cambridge MA: The MIT

Press).

DeAngelis, G.C., and Uka, T. (2003). Coding of horizontal disparity and velocity by MT neurons in the alert macaque. J. Neurophysiol. *89*, 1094–1111.

Desimone, R., and Schein, S.J. (1987). Visual properties of neurons in area V4 of the macaque: sensitivity to stimulus form. J. Neurophysiol. *57*, 835–868.

Dorris, M.C., and Glimcher, P.W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron *44*, 365–378.

Dorris, M.C., Olivier, E., and Munoz, D.P. (2007). Competitive integration of visual and preparatory signals in the superior colliculus during saccadic programming. J. Neurosci. *27*, 5053–5062.

Elston, G.N., and Rosa, M.G. (1997). The occipitoparietal pathway of the macaque monkey: comparison of pyramidal cell morphology in layer III of functionally related cortical visual areas. Cereb. Cortex *7*, 432–452.

Falkner, A.L., Krishna, B.S., and Goldberg, M.E. (2010). Surround suppression sharpens the priority map in the lateral intraparietal area. J. Neurosci. *30*, 12787–12797.

Ferrera, V.P., Yanike, M., and Cassanello, C. (2009). Frontal eye field neurons signal changes in decision criteria. Nat. Neurosci. *12*, 1458–1462.

Fitzgerald, J.K., Freedman, D.J., Fanini, A., Bennur, S., Gold, J.I., and Assad, J.A. (2013). Biased associative representations in parietal cortex. Neuron *77*, 180–191.

Freedman, D.J., and Assad, J.A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature *443*, 85–88.

Freedman, D.J., and Assad, J.A. (2011). A proposed common neural mechanism for categorization and perceptual decisions. Nat. Neurosci. *14*, 143–146.

Ganguli, S., Bisley, J.W., Roitman, J.D., Shadlen, M.N., Goldberg, M.E., and Miller, K.D. (2008). One-dimensional dynamics of attention and decision making in LIP. Neuron *58*, 15–25.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. Annu. Rev. Neurosci. *30*, 535–574.

Goldman, M.S. (2009). Memory without feedback in a neural network. Neuron *61*, 621–634.

Gottlieb, J. (2007). From thought to action: the parietal cortex as a bridge between perception, action, and cognition. Neuron *53*, 9–16.

Gottlieb, J., Balan, P., Oristaglio, J., and Suzuki, M. (2009). Parietal control of attentional guidance: the significance of sensory, motivational and motor factors. Neurobiol. Learn. Mem. *91*, 121–128.

Hegdé, J., and Van Essen, D.C. (2007). A comparative study of shape representation in macaque visual areas V2 and V4. Cereb. Cortex *17*, 1100–1116.

Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol. *160*, 106–154.2.

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. Nat. Rev. Neurosci. *2*, 194–203.

Kable, J.W., and Glimcher, P.W. (2009). The neurobiology of decision: consensus and controversy. Neuron *63*, 733–745.

Kennerley, S.W., and Wallis, J.D. (2009). Reward-dependent modulation of working memory in lateral prefrontal cortex. J. Neurosci. *29*, 3259–3270.

Lehky, S.R., Kiani, R., Esteky, H., and Tanaka, K. (2011). Statistics of visual responses

in primate inferotemporal cortex to object stimuli. J. Neurophysiol. *106*, 1097–1117.

Leon, M.I., and Shadlen, M.N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron *24*, 415–425.

Lewis, J.W., and Van Essen, D.C. (2000). Corticocortical connections of visual, sensorimotor, and multimodal processing areas in the parietal lobe of the macaque monkey. J. Comp. Neurol. *428*, 112–137.

LoFaro, T., Louie, K., Webb, R., and Glimcher, P.W. (2014). The Temporal Dynamics of Cortical Normalization Models of Decision-making. Lett. Biomath. *1*, 209–220.

Louie, K., Grattan, L.E., and Glimcher, P.W. (2011). Reward value-based gain control: divisive normalization in parietal cortex. J. Neurosci. *31*, 10627–10639.

Louie, K., LoFaro, T., Webb, R., and Glimcher, P.W. (2014). Dynamic Divisive Normalization Predicts Time-Varying Value Coding in Decision-Related Circuits. J. Neurosci. *34*, 16046–16057.

Markov, N.T., Misery, P., Falchier, A., Lamy, C., Vezoli, J., Quilodran, R., Gariel, M.A., and Giroud, P. (2011). Weight Consistency Specifies Regularities of Macaque Cortical Networks. Cereb. Cortex *21*, 1254–1272.

Mazurek, M.E., Roitman, J.D., Ditterich, J., and Shadlen, M.N. (2003). A Role for Neural Integrators in Perceptual Decision Making. Cereb. Cortex *13*, 1257–1269.

Miller, E.K., and Wilson, M.A. (2008). All my circuits: using multiple electrodes to understand functioning neural networks. Neuron *60*, 483–488.

Murphy, B.K., and Miller, K.D. (2009). Balanced amplification: a new mechanism of

selective amplification of neural activity patterns. Neuron *61*, 635–648.

Nurminen, L., and Angelucci, A. (2014). Multiple components of surround modulation in primary visual cortex: Multiple neural circuits with multiple functions? Vision Res. *104*, 47–56.

Oristaglio, J., Schneider, D.M., Balan, P.F., and Gottlieb, J. (2006). Integration of visuospatial and effector information during symbolically cued limb movements in monkey lateral intraparietal area. J. Neurosci. *26*, 8310–8319.

Ozeki, H., Finn, I.M., Schaffer, E.S., Miller, K.D., and Ferster, D. (2009). Inhibitory stabilization of the cortical network underlies visual surround suppression. Neuron *62*, 578–592.

Pan, X., Sawa, K., Tsuda, I., Tsukada, M., and Sakagami, M. (2008). Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. Nat. Neurosci. *11*, 703–712.

Patel, G.H., Shulman, G.L., Baker, J.T., Akbudak, E., Snyder, A.Z., Snyder, L.H., and Corbetta, M. (2010). Topographic organization of macaque area LIP. Proc. Natl. Acad. Sci. U. S. A. *107*, 4728–4733.

Peck, C.J., Jangraw, D.C., Suzuki, M., Efem, R., and Gottlieb, J. (2009). Reward Modulates Attention Independently of Action Value in Posterior Parietal Cortex. J. Neurosci. *29*, 11182–11191.

Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.

Roesch, M.R., and Olson, C.R. (2003). Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. J.

Neurophysiol. *90*, 1766–1789.

Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. *22*, 9475–9489.

Rubin, D.B., Hooser, S.D. Van, and Miller, K.D. (2015). The stabilized supralinear network : A unifying circuit motif underlying multi-input integration in sensory cortex. Neuron *85*, 402-417.

Schall, J.D., and Hanes, D.P. (1993). Neural basis of saccade target selection in frontal eye field during visual search. Nature *366*, 467–469.

Schall, J.D., Morel, A., King, D.J., and Bullier, J. (1995). Topography of visual cortex connections with frontal eye field in macaque: convergence and segregation of processing streams. J. Neurosci. *15*, 4464–4487.

Schein, S.J., and Desimone, R. (1990). Spectral properties of V4 neurons in the macaque. J. Neurosci. *10*, 3369–3389.

Shenoy, K. V, Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. Annu. Rev. Neurosci. *36*, 337–359.

Stanton, G.B., Bruce, C.J., and Goldberg, M.E. (1995). Topography of projections to posterior cortical areas from the macaque frontal eye fields. J. Comp. Neurol. *353*, 291–305.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching Behavior and the representaiton of value in the parietal cortex. Science *304*, 457–461.

Sundberg, K. a, Mitchell, J.F., and Reynolds, J.H. (2009). Spatial attention modulates center-surround interactions in macaque visual area v4. Neuron *61*, 952–963.

Suzuki, M., and Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. Nat. Neurosci. *16*, 98–104.

Swaminathan, S.K., and Freedman, D.J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat. Neurosci. *15*, 315–320.

Teichert, T., Yu, D., and Ferrera, V.P. (2014). Performance monitoring in monkey frontal eye field. J. Neurosci. *34*, 1657–1671.

Tsui, J.M.G., and Pack, C.C. (2011). Contrast sensitivity of MT receptive field centers and surrounds. J. Neurophysiol. *106*, 1888–1900.

Usher, M., and McClelland, J.L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychol. Rev. *108*, 550–592.

Wang, X.J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron *36*, 955–968.

Wong, K.-F., and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. *26*, 1314–1328.

Xing, J., and Andersen, R. A. (2000). Memory activity of LIP neurons for sequential eye movements simulated with neural networks. J. Neurophysiol. *84*, 651–665.